EFFICIENCY ASPECTS OF DESIGN AND ANALYSIS OF PROSPECTIVE COHORT STUDIES ON DIET, NUTRITION AND CANCER.

Rudolf J. Kaaks

Promotoren : mw. Dr. W.A. van Staveren
Bijzonder hoogleraar in de voeding van de oudere mens

Dr. J.G.A.J. Hautvast
Hoogleraar in de leer van de voeding en de voedselbereiding

Co-promotor : Elio Riboli, M.D., M.Sc.
Head Nutrition and Cancer programme,
International Agency for Research on Cancer, Lyon, France.
Adjunct professor in environmental medicine,
New York University Medical Centre.

EFFICIENCY ASPECTS OF DESIGN AND ANALYSIS OF PROSPECTIVE COHORT STUDIES ON DIET, NUTRITION AND CANCER.

Rudolf J. Kaaks

Proefschrift

ter verkrijging van de graad van doctor
in de landbouw- en milieuwetenschappen
op gezag van de rector magnificus
dr. C.M. Karssen,
in het openbaar te verdedigen
op vrijdag 7 oktober 1994
des namiddags om half twee in de Aula
van de Landbouwuniversiteit te Wageningen

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Kaaks, Rudolf Johan
Efficiency aspects of design and analysis of prospective cohort studies on diet, nutrition and cancer / Rudolf J. Kaaks.- [S.l.:s.n.]
Thesis Wageningen. - With ref. - With summary in Dutch, and French.
ISBN 90-5485-314-X

Cover design : Ernst van Cleef
Printing : Grafisch Service Centrum Van Gils B.V.
Wageningen, The Netherlands

STELLINGEN

1. "Validatie" en "calibratie" van voedingsinname-metingen moeten per definitie worden beschouwd als vormen van statistische modellering, berustend op de aanname dat verschillende methoden eenzelfde verschijnsel meten, maar onafhankelijke fouten hebben.
(dit proefschrift)

2. Validiteit en precisie kunnen het best worden gezien als kenmerken van een verzameling metingen, en niet van de gebruikte meetmethode.
(dit proefschrift)

3. Bland & Altman's categorische verwerping (Lancet,1986;i:307-10) van de correlatie-coefficient als maat voor validiteit of reproduceerbaarheid (of een combinatie van beide) is onterecht.

4. Het onderzoeken van epidemiologische verbanden tussen innameniveaus van individuele nutrienten of voedingsmiddelen en chronische ziekten leidt gemakkelijk tot overinterpretaties die zouden kunnen worden vermeden als de voedings-"blootstelling" meer integraal werd beschreven als een multi-dimensionaal consumptiepatroon.

5. De recente uitkomsten van de finse kanker-preventie trial met β-caroteen en α-tocopherol (New England Journal of Medicine, 1994;330:1029-35) suggereren dat men voor chemopreventie beter wat vaker naar de groenteboer kan gaan dan naar de apotheek.

6. De recent ingevoerde registratie van epidemiologen in Nederland suggereert een gebrek aan vertrouwen dat (collega) werkgevers van epidemiologisch onderzoekers zelf een curriculum vitae kunnen beoordelen.

7. De sociale rechtvaardigheid van een wettelijk gegarandeerd minimum inkomen is aangetast als men werkzoekenden niet op afzienbare termijn een baan kan bieden.

8. De sterke toename van aantal skiliften in franse Alpen doet vermoeden dat, onder het motto "vooruitgang is alleen vooruitgang als die wordt gedeeld door iedereen", de wereld langzaam wordt omgetoverd in een gigantisch gemechaniseerd pretpark.

9. Des chercheurs qui cherchent, on en trouve; des chercheurs qui trouvent, on en cherche. (generaal de Gaulle)

10. Te beoordelen naar hun gedrag in het stadsverkeer hebben de fransen met de haan als nationaal symbool geen slechte keus gemaakt.

11. De aankondiging van een verhoging van het budget voor de vroege opsporing van mammacarcinoom (Volkskrant, 1-9-94) valt op aardige wijze samen met de recente benoeming van mevrouw Borst als minister voor Volksgezondheid, Welzijn en Sport.

Stellingen behorend bij het proefschrift 'Efficiency aspects of design and analysis of prospective studies on diet, nutrition and cancer', van Rudolf Kaaks. Wageningen, 7 oktober 1994.

Abstract

This thesis presents and analyzes methodological approaches to improve the design and analysis of prospective cohort studies on the relations between diet, nutritional status and cancer. The first chapters discuss methods to optimize the measurement of the individuals' habitual dietary intakes, focussing on the use and design of sub-studies for the "validation" or "calibration" of baseline dietary questionnaire assessments. The power of prospective studies can be improved by maximizing the variation in true dietary intake levels actually distinguished – or "predicted" – by questionnaire assessments. This can be achieved by designing an optimal questionnaire method, using a preliminary validity study to evaluate its performance. An additional possibility is to broaden the range of dietary exposures by conducting multiple cohort studies in populations with different dietary habits. A main objective is to precisely estimate the magnitude of the predicted variation of intake levels, to account for the effect of measurement error as well as of the real variation in exposure, in the evaluation of the power or sample size requirements of a cohort study, and in the estimation of relative risks describing diet-disease relations. The predicted variation is estimated most efficiently by means of a "calibration" sub-study, which differs from validity studies in that it requires only a single (unbiased) reference measurement per person (e.g., based on a 24-hour recall), in a representative sub-sample of cohort members. In multi-cohort projects, calibration studies are essential to improve the between-cohort comparability of relative risk estimates, and to increase the power of a statistical test for the presence of a diet-disease association based on a pooled summary estimate. A simplified method is proposed for the estimation of sample size requirements of dietary calibration studies. When the exposure assessments are based on a biochemical marker, a most efficient design is to store biological specimens in a biobank, and to postpone laboratory analyses until cases with disease have been identified. Nevertheless, the number of scientific hypotheses potentially of interest is usually much larger than can be tested with limited amounts of biological specimens available. The last chapter of this thesis discusses the use of a sequential study design, to allow the evaluation of a maximum number of different hypotheses at the expense of as little biological material as possible.

The research described in this thesis was carried out at the International Agency for Research on Cancer (Lyon, France) as part of the EPIC project, supported by a grant from the Europe Against Cancer Programme of the Commission of European Communities

CONTENTS

Chapter 1

General introduction

General introduction

Background

Scope of this thesis

i. "Validation" and "calibration" of dietary intake assessments

ii. Design aspects when exposure measurements are based on biochemical markers

Background

Over the past 20 or 30 years, important developments have been made in epidemiological research on the relation between diet and cancer. Following international correlation studies (1-3), and studies on migrants (4,5), which indicated that diet and nutrition-related life style may be important determinants of cancer risk, epidemiological research shifted towards studies where the basic units of observation were individuals rather than entire populations. During the 1970s, most of these were of a case-control design, focusing mainly on cancers of the stomach, colorectum, and breast. During the 1980s, the number of case-control studies increased, and were gradually oriented to a larger variety of cancer sites, including the upper aerodigestive tract (larynx, oesophagus), endometrium, ovary, prostate and lung (6). This period was also characterized by the development of more modern concepts and methods for "nutritional epidemiology", as evidenced by the (still relatively recent) publication of two standard textbooks in this field (7,8). Particular attention was given to the development of appropriate methods for the assessment of individuals' habitual diet, especially food frequency questionnaires (7,9), and to the use of methodological sub-studies to evaluate the validity and reproducibility of the dietary questionnaire assessments (10,11). Finally, a number of large, well-designed prospective cohort studies were started, in which diet was measured by means of carefully selected and "validated", questionnaire instruments (12-14). Then, at the end of the 80s, the idea of developing multi-cohort projects was conceived. The first of these was the European Prospective Investigation on Cancer and Nutrition (EPIC), a multi-centre cohort study currently being conducted in collaboration with 17 research centres in seven European countries, and which is coordinated by the International Agency for Research on Cancer at Lyon (15). Following a similar rationale, another multi-centre (and multi-ethnic) project is being planned in areas around Pacific basin, including Hawaii, Califoria, Singapore and possibly Japan (16).

The main reasons for conducting prospective cohort studies, rather than using a case-control design, are that in the latter the estimated association between diet and disease risk may be prone to bias. Selection biases may occur if controls and cases do not originate from the same population base (17). Another type of bias is due to differences between

cases and controls in the recall of their previous dietary habits (18). In prospective studies, the diet-disease relationship is investigated following the natural time sequence between the exposure, assessed at baseline when the subjects enrol in the study, and the subsequent occurrence of disease during a period of follow up. It is unlikely that assessments of habitual dietary intake will be differently biased among participants who eventually develop a given disease, as compared to those who remain in good health. Moreover, prospective cohort studies provide a well described population base for making comparisons between the measured exposures of cases, and of disease-free control subjects. Thus, selection bias is also unlikely to form a serious problem in this type of study, unless for some reason losses in follow up are associated with the level of dietary intake assessments.

An important remaining problem, also in prospective cohort studies, is that there should be sufficient statistical power to test for the presence (or absence) of specific diet-disease associations. In conjunction, it is desirable to obtain relative risk estimates with a sufficient level of precision, as measured by the width of their confidence intervals. Within a limited geographical (or cultural) area, there may be relatively little between-individual variation in habitual dietary intake, as compared to the variation that exists between mean intake levels in different countries. For example, the mean fat intake at a population level varies from as little as 11 percent of daily energy intake for some developing countries, to more than 43 percent in the United States, whereas within the United States as a single country the between-individual variation in fat intake was estimated to be between 30 and 45 percent of total energy (19). Due to this relative homogeneity of dietary habits within a single country, true relative risks between subjects with either "high" or "low" intake levels of a given food or nutrient will tend to be much weaker. In addition, there is the problem that relative risks tend to be under-estimated due to the attenuating effects of random errors in the dietary exposure measurements, so that the statistical power to test for the presence of a diet-disease association is even further decreased (20). Therefore, relatively large numbers of cases with the disease of interest will be required for a cohort study to reach a reasonable level of power and precision.

Although the total cancer burden is high in economically more developed countries, there are many different forms of cancer, and incidence rates of each of these separately are usually relatively low. Therefore, prospective

cohort studies must in general be very large, including several tens of thousands of individuals, for a sufficient number of cases with a specific form of disease to develop during a follow up period of no more than 10 to 15 years. Consequently, the costs of such prospective studies can be very high. It is thus fundamental to use an efficient study design which, for a given investment of time and resources, makes the study as informative as possible. The main criteria to judge the amount of information obtained in a study are:

a. the power of statistical tests for the presence of specific diet-disease associations;
b. the validity and precision with which the magnitude of such associations can be estimated (e.g., in the form of relative risks); and
c. the number of different scientific hypotheses that can be evaluated.

Major aspects of the design of a prospective study are related to the choice of the study population, as characterized by the expected cumulative incidence of disease (within a given follow-up period), the presence of a reliable mechanism for follow-up (e.g., a cancer registry), or factors which may facilitate contacting the study subjects (e.g., participation in a local screening programme). Once a choice has been made for the type of study population, the efficiency of the design of a prospective cohort study can be optimized by maximizing the accuracy of the exposure measurements, and by determining the sample size at which the cohort will have sufficient statistical power. The estimated sample size requirements, as well as financial resources available, are then key elements for deciding how many different types of exposure information can be collected from each participant (e.g., apart from the main questionnaire(s) on dietary habits, additional questionnaires can be included, for instance on physical activity, or biological samples can be collected for the assessment of various biochemical markers).

Scope of this thesis

This thesis addresses a number of methodological issues related to the efficiency of the design and conduct of prospective cohort studies. The first chapters (chapters 2 to 5) discuss methods for optimizing the assessment of the habitual, long-term dietary intake of individuals participating in a prospective cohort study. More specifically, these chapters focus on the use and design of sub-studies with additional reference measurements, for "validation" or "calibration" of baseline questionnaire assessments of dietary intake level. Chapter 6, on the other hand, addresses the aspect of optimizing the number of specific study hypotheses that can be evaluated when exposure assessments are based on a biochemical marker measured in blood, or other tissue samples.

i. "Validation" and "calibration" of dietary intake assessments

As mentioned above, the power and precision of a cohort study on diet depend, among other things, on the heterogeneity in dietary intake levels within a given study population, as well as on the accuracy with which this variation in intake level is measured at baseline. Therefore, before starting the main epidemiological study, it is important to verify whether the baseline dietary intake assessments – usually obtained by means of a structured food frequency, or dietary history type of questionnaire (9,10) – make sufficient distinctions between the high or low intake levels of different individuals. For this purpose, it is usually proposed to conduct a smaller sub-study, in which the accuracy of questionnaire assessments is evaluated by comparison with 'reference' measurements that are assumed to provide a more accurate measure of the individuals' true habitual intake levels (7). In the first instance, such sub-studies can be used during the development (21) or selection (22-24) of an optimal dietary questionnaire instrument, even before the main cohort study is started. Chapter 2 of this thesis presents a mathematical model for the definition of different types of error in dietary exposure measurements. This chapter then reviews, in terms of latent variable models, the essential requirements for the design and analysis of dietary validity studies, aimed at estimating the correlation between questionnaire assessments and the true, habitual dietary intake levels of individuals.

Apart from selecting an optimal questionnaire instrument, further reasons for collecting additional reference measurements are that:

1. at the start of a prospective cohort study, this allows a more precise estimation of the expected statistical power, or sample size requirements of the cohort, taking account of the inaccuracy of certain dietary intake assessments; and
2. at the analysis stage, this will allow the estimation of relative risk estimates with a correction for biases due to errors in the baseline dietary exposure assessments. Indeed, we are interested in disease risk as a function of true dietary intake levels rather than of measured levels of intake. With only a single, baseline measurement of dietary exposure (usually obtained by means of a questionnaire), the quantitative relation between true intake level and disease risk cannot be estimated.

In previous epidemiological studies, the additional reference measurements needed to meet these two objectives have been usually collected within a preliminary validity study. Prospective cohort studies, however, also offer the possibility of collecting reference measurements as an integral part of the overall dietary exposure assessment at baseline, on at least a representative sample of study participants. This possibility is discussed in Chapter 3, which proposes an efficient alternative to dietary validity studies, based on the concept of "calibration" of the baseline dietary questionnaire assessments.

Another possible approach to improve the correlation between measured and true dietary exposure values, and to increase the power and precision of a prospective cohort study while keeping its sample size constant, is to broaden the range of true dietary exposure levels covered. This may be achieved by conducting studies in different geographical areas (as in the EPIC project (15)), or by including different ethnic sub-groups which are known to have different food consumption habits (this is the rationale for the Pacific Area Multi-Ethnic project (16)). Advantages of the multi-cohort study design, as compared to "ecological" studies based on aggregate (i.e., group level) information about exposure and disease incidence, are discussed in Chapter 4. A complication in the multi-cohort design, however, is that, within different cohorts, questionnaire assessments of dietary exposure may not have the same degree of accuracy for classification of individuals by

their habitual dietary intake levels, whereas mean intake levels may also be over- or underestimated by unequal amounts. A solution to this problem is presented in Chapter 4, proposing the "calibration" approach to combine the findings of different cohorts in a manner that reflects the accuracy of the questionnaire assessments, and to adjust for differences in systematic over- or underestimation of mean intakes at a level.

Within this context, it is important that calibration sub-studies should themselves be large enough to reach a minimum level of precision, without on the other hand overinvesting in this component of exposure assessment. The issue of optimal sample size requirements for dietary calibration studies is addressed in Chapter 5.

ii. Design aspects when exposure measurements are based on biochemical markers

Another important advantage of prospective cohort studies is the possibility to use exposure assessments based on biochemical markers, measured in urine or blood, or in tissue specimens such as nails and fat tissue biopsies. Since in prospective studies the biological specimen can be collected before the clinical manifestation of disease, it is unlikely that the presence of a tumour, or related metabolic effects such as cancer cachexia, will have influenced the levels of biochemical marker. Observed associations between biochemical markers and disease risk can thus be interpreted more reliably as reflecting a causal relation, between the level of a given type of exposure and the development of a disease (and not vice versa). It should not be forgotten, in the current development of "biochemical" epidemiology, that the time sequence between exposure and disease is one of the fundamental conditions for interpretation of an epidemiological association as a potentially causal one (25). Markers which of may be of interest in studies on diet and cancer include:

- markers of dietary intake and nutritional status (e.g. plasma levels of vitamins, triglycerides, or lipoproteins; fatty acid composition of fat tissue biopsies, selenium levels in toenails) (26-28),
- markers of hormonal status and metabolism (e.g., plasma levels of specific steroid hormones, or sex-hormone binding globulin) (29),
- markers of susceptibility (e.g., genetic or phenotypic polymorphisms of enzymes which may play a role in the activation or inactivation of (pre-) carcinogens (30)),

- markers of DNA damage (e.g., oxidative damage of DNA (31,32).

Efficient approaches to exposure measurements by means of biochemical markers differ from those where exposure assessments are obtained by means of a questionnaire, because:

1. it would be too expensive to measure all biochemical markers potentially of interest at baseline for all individuals; and
2. the types of marker of interest will vary according to the prevailing biological hypotheses for the type of cancer under investigation.

It is therefore usually more efficient to store biological specimens in a biological bank, and to delay the second step of the exposure assessment (the laboratory analysis) until a later date, when it will be known which individuals have developed a given type of disease, and which will be selected as suitable control subjects. Nevertheless, there remains the problem that the number of biological hypotheses potentially of interest will generally exceed the number of biochemical parameters that can actually be assessed with the limited volume of biological specimens for cases and suitable controls. For instance, in the EPIC project (15), a total of only 14 millilitres of blood fractions (plasma, serum, buffy coat and red blood cells) are kept for each individual, in the form of 28 smaller (0.5 ml.) aliquots. Therefore, after the creation of a biological bank, an additional aspect related to increasing the efficiency of exposure assessment is how to optimize the number of different hypotheses that can be evaluated with a limited biological material. This aspect is addressed in Chapter 6, which proposes the use of a sequential test procedure to distinguish between promising new hypotheses, which may be worth further investigation, and less promising ones.

References

1. Armstrong BK, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. Int J Cancer 1975; 15:617-31.
2. Drasar BS, Irving. Environmental factors and cancer of the colon and breast. Br J Cancer 1973; 27:167-72
3. Howell MA. Factor analysis of international cancer mortality data and per capita food consumption. Br J Cancer 1974; 29:328-36.
4. Dunn JE. Breast cancer among American Japanese in the San Francisco Bay area. Natl Cancer Inst Monogr 1977; 32:73-79.

5. McMichael AJ, McCall MG, Hartshorne JM, Woodings TL. Patterns of gastro-intestinal cancer in European migrants to Australia. The role of dietary change. Int J Cancer 1980; 25: 431-7.
6. Tomatis L, Aitio A, Day NE, Heseltine E, Kaldor J, Miller AB, Parkin DM, Riboli E. Cancer: causes, occurrence and control. (IARC scientific publications No. 100). Lyon, IARC (1990).
7. Willett W. Nutritional epidemiology. Oxford University Press. New York, 1990. Chapter 6: Reproducibility and validity of food frequency questionnaires.
8. Margetts BM, Nelson M (eds). Design concepts in nutritional epidemiology. Oxford, Oxford University Press (1991).
9. Cameron ME, van Staveren WA (eds). Manual on methodology for food consumption studies. Oxford, Oxford University Press (1988).
10. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE. Reproducibility and validity of a semi-quantitative food frequency questionnaire. Am J Epidemiol 1985; 122:51-65.
11. van Staveren WA, de Boer JO, Burema J. Validity and reproducibility of a dietary history method estimating the usual food intake during one month. Am J Clin Nutr 1985; 42:554-9.
12. Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Hennekens CH, Speizer FE. Dietary fat and the risk of breast cancer. N Engl J Med 1987; 316:22-8.
13. Howe GR, Friedenreich CM, Jain M, Miller AB. A cohort study of fat intake and risk of breast cancer. J Natl Cancer Inst 1991; 83:336-40.
14. van den Brandt PA, Goldbohm RA, van 't Veer P, Volovics A, Hermus RJJ, Sturmans F. A large-scale prospective cohort study on diet and cancer in the Netherlands. J Clin Epidemiol 1990; 43:285-95.
15. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer an Nutrition (EPIC). Ann oncol 1992; 3:783-91.
16. Dr. L. Kolonel. Epidemiology programme, University of Hawaii, United States. Personal communication.
17. Schlesselman JJ. Case-control studies: design, conduct, analysis. New York, Oxford University Press (1982).
18. Giovanucci E, Stampfer MJ, Colditz GA, Manson J, Rosner B, Longnecker M. A comparison of prospective and retrospective assessments of diet in the study of breast cancer. Am J Epidemiol 1993; 137:502-11.
19. Wynder EL, Hebert JR. Homogeneity and nutritional exposure: an impediment in cancer epidemiology? J Natl Cancer Inst 1987; 79:605-7.
20. Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiological studies of diet and cancer. Nutr Cancer 1988; 11:243-50.
21. Jain MG, Harrison L, Howe GR, Miller AB. Evaluation of a self-administered dietary questionnaire for use in a cohort study. Am J Clin Nutr 1982; 36:931-5
22. Pietinen P, Hartman AM, Haapa E, Räsänen L, Haapakoski J, Palmgren J, Albanes D, Virtamo J, Huttunen JK. Reproducibility and validity of dietary assessment instruments. I Self-administered food use questionnaire with portion size picture booklet. Am J Epidemiol 1988; 128:655-66.
23. Pietinen P, Hartman AM, Haapa E, Räsänen L, Haapakoski J, Palmgren J, Albanes D, Virtamo J, Huttunen JK. Reproducibility and validity of dietary assessment instruments. II A qualitative food frequency questionnaire. Am J Epidemiol 1988; 128:655-66.

24. Riboli E, Elmsthal S, Saracci R, Gullberg B, Lindgärde F. The Malmö food study: validity of two dietary assessment methods for measuring nutrient intake. Am J Epidemiol (submitted)
25. Hill AB. The environment and disease: association or causation? Proc R Soc Med 58, 295-300.
26. Riboli E, Rönnholm H, Saracci R. Biological markers of diet. Cancer Surveys 1987; 6:685-718
27. Hunter D. Biochemical indicators of dietary intake. Guest chapter in Willett W. Nutritional epidemiology, Oxford University Press, New York, 1990.
28. Kok F, van 't Veer P. (eds.) Biomarkers of dietary exposure. Proceedings of the 3rd Meeting onNutritional Epidemiology, Smith-Gordon, London, 1991.
29. Key TJA, Beral V. Sex hormones and cancer. In: Vanio H, Magee P, McGregor D, McMichael AJ (eds.): Mechanisms of carcinogenesis in risk identification. (IARC Scientific Publications No. 116). Lyon, IARC (1992).
30. Bartsch H, Armstrong BK. Host factors in human carcinogenesis. (IARC Scientific Publications No. 39). Lyon, IARC (1982).
31. Bartsch H, Hemminki K, O'Neill IK. Methods for detecting DNA damaging agents in humans: applications in cancer epidemiology and prevention. (IARC Scientific Publications No. 89). Lyon, IARC (1988).
32. Simic MG. DNA markers of oxidative processes in vivo: relevance to carinogenesis and anticarcinogenesis. Cancer Res 1994; 54:1918-23.

Chapter 2

Estimating the accuracy of of dietary questionnaire assessments: Validation in terms of structural equation models.

This chapter has been published by the authors R. Kaaks, E. Riboli, J. Estève, A.L. van Kappel, and W.A. van Staveren, in Statistics in Medicine, 1994;13:127-42.

Abstract

The validity and precision of questionnaire assessments of the habitual intake of individuals are usually evaluated by comparison with reference measurements that are supposed to provide a best possible substitute for the individuals' true intake values. In the present paper, a measurement error model is presented, defining different types of error - random or systematic, and within or between individuals - that may occur in dietary intake measurements. It is then discussed how simple latent variable models (structural equation models) can be used to estimate the average magnitude of these various types of error. So far, approaches described for the analysis of dietary validity studies were all based on the assumption that the random errors of repeat reference measurements, taken by the same method on different occasions, are uncorrelated, so that the average of a sufficiently large number of repeat reference measurements will provide an accurate ranking of individuals by true intake level. In the present paper it is described how, by additional comparison with a third type of measurement such as a biochemical marker, the validity of dietary questionnaire measurements can be evaluated even in situations where the random errors of repeat reference measurements cannot be assumed to be independent.

Introduction

The validity and precision of measurements of diet obtained with dietary questionnaires are usually studied by evaluating their concordance with "reference" measurements, which are supposed to provide the best possible substitute for the true habitual intake value of every single subject participating in a dietary validity study (1-3). The two classic measures of concordance most commonly used in such validity studies are the correlation coefficient, and the difference between group means. The first measure is seen as an index of the accuracy with which questionnaire assessments can rank individuals by dietary intake level, while the second is used to express the average tendency of individuals to over- or under-estimate their dietary intake.

In most studies on the validity of dietary questionnaire assessments, the reference measurements have been based on weighed food records (2), but 24-hour recalls have also been used (2,4). Usually, such measurements of daily intake are repeated for multiple days, with the objective of obtaining a precise estimate of the individual level of the usual daily nutrient intake value. In early validity studies it was often assumed that the average of one or two weeks of daily intake recording provided accurate measurement of long-term nutrient intake. On this assumption, the correlation between questionnaire assessments and reference measurements should precisely reflect the accuracy of the questionnaire assessments for ranking individuals by habitual intake level. More recently it was shown that a larger number of recording days may be required, and that the correlation of questionnaire with reference measurements might underestimate the correlation with true intake levels if the reference measurements are subject to random error (e.g., if the reference measurements are based on only one week of food records) (5-8). Methods were proposed to correct for the underestimation of correlation and regresssion coefficients ("attenuation" bias) due to within-subject random errors in the reference measurements (6,9). Freedman and colleagues (10) extended these methods by also incorporating an adjustment for the over- or under-estimation of correlation and regression coefficients that may result from a covariance between random errors of questionnaire and reference measurements taken too close in time.

So far, the various approaches described for the analysis of dietary validity studies were all based on the assumption that the average of repeat

reference measurements will provide an accurate ranking by true intake level provided that the number of repeat measurements is large enough. In practice, however, this may not be the case. For instance, by comparing weighed food records with precise measurements of energy expenditure obtained by the "doubly labeled water" method (11,12), it has been shown that individuals can differ in their tendency to systematically under-report energy intake. One may expect that 24-hour recalls would result in an even greater systematic under-reporting, since a major source of error will be the subjects' tendency to forget which foods they actually consumed during the previous day. In addition, one would expect that greater systematic errors will be made in the description of portion sizes (2). Individuals might therefore systematically differ in their over- or under-reporting of intake. If this is the case, the random errors of repeat reference measurements, taken by the same method on different occasions, will be correlated, and repetition of daily intake estimates will fail to provide a fully accurate classification of individuals by true intake level.

In the present paper, we shall present a measurement error model, defining different types of error - random and systematic, within or between individuals (13) - that may occur in dietary intake measurements. We shall then review, in terms of structural equation models, various procedures for the estimation of the average magnitude of the different types of error, and show how simple SAS programmes (using the procedure "CALIS") can be used for computations. Basic assumptions underlying the analysis of dietary validity studies will be discussed. It will be shown that, by additional comparison with a third type of measurement such as a biochemical marker, unbiased estimates of the average magnitude of the different errors can be obtained even when the random errors of repeat reference measurements are correlated.

A measurement error model

Suppose that for different individuals within a dietary study one seeks to measure the usual intake of a given nutrient. If, for a given individual i, a measurement X is made of his or her true intake T, one can write :

$$X_i = T_i + \gamma_i$$
$$E(\gamma_i) = B_i$$
$$Var(\gamma_i) = \sigma^2_{\gamma,i}$$

where γ_i is the "error", B_i is the bias, and $\sigma^2_{\gamma,i}$ is the variance of the within-subject random error (13), which is the random error for repeat

measurements taken by the same method on the same individual. B_i has previously been called the within-subject systematic error (13), or subject-specific bias (2,3), and differs from zero if, on repeat occasions, the individual tends consistently to over- or underestimate intake.

Within a population of I individuals, the magnitude of the bias B_i and of the variance $\sigma^2_{\gamma,i}$ may vary from one subject to another. On average, however, the variance of within-subject random error will be equal to σ^2_{γ}, the expected individual variance in the population. The average variance σ^2_{γ} will be assumed to be the same at all values of true intake, T.

The bias may, to a certain extent, be functionally related to T. We shall assume that this functional relation is linear:

$$B = \alpha + bT + \delta$$

where $E(\delta) = 0$, $Var(\delta) = \sigma^2_{\delta}$, and $Cov(\delta,T) = 0$.

According to this linear model, the expected over- or underestimation for each individual can be decomposed into a multiplicative component bT, which depends on the individual's true dietary intake, and a constant additive component α. The term δ represents the residual part of the subject-specific bias, which cannot be accurately predicted from the linear relation with T. We shall therefore refer to δ as "random" bias. Also the variance σ^2_{δ} will be assumed to be equal for all values of T.

We can now write the following model to measure the intake for an individual i belonging to the given study population :

$$X_i = T_i + (\alpha + bT_i) + \delta_i + \gamma_i = \alpha + \beta T_i + \varepsilon_i ,$$

Here, the measurement X_i is the sum of the expected measurement, conditional on the individual's true intake,

$$E(X|T_i) = T_i + (\alpha + bT_i) = \alpha + \beta T_i ,$$

plus a <u>total</u> random error ε_i. The total random error is itself the sum of γ_i, the within-subject random error, and of δ_i, the unpredictable part of the individual's bias. The distinction between the total random error components γ and δ is of importance since, when X measurements are repeated on the same individuals, the random biases δ_i will be reproduced. As a consequence, the total random errors ε of the repeat measurements will be correlated unless all $\delta_i = 0$. The average variance of the total random error is given by

$$Var(\varepsilon) = Var(X|T) = \sigma^2_{\delta} + \sigma^2_{\gamma} = \sigma^2_{\varepsilon} .$$

The coefficient α indicates the average tendency to over- or under-estimate intake by a constant amount, and β indicates the tendency to over- or under-

estimate intake by an amount which is proportional to the level of the true intake.

"Validity" is commonly defined as the absence of bias, whereas "precision" is usually defined to be equivalent to a high reproducibility of measurements (2). Thus, if the objective is to measure the intake of individuals, validity corresponds in model terms with the absence not only of constant and proportional biases (i.e., α=0, β=1.0), but also of random biases (i.e., σ^2_δ=0). A high reproducibility then corresponds to a small variance of within-subject random errors. It should be kept in mind that the parameters in the estimated measurement model (α, β, σ^2_δ, and σ^2_γ) may not only depend on the type of measuring instrument but also on the population being investigated, and therefore should not be taken as universal parameters specific for the measuring instrument used.

If the values of the parameters α and β were known, one could correct the X-measurements for "systematic" biases by: 1) substracting an amount α to correct for the average additive bias, and 2) subsequent division by β to correct for the average multiplicative bias. The relation between the corrected measurements and true intake then becomes:

$X_i^* = (X_i-\alpha)/\beta = T_i + \varepsilon_i^*$, where $\varepsilon_i^* = \varepsilon_i/\beta$.

This correction can be seen as a scale adjustment, re-expressing the X-measurements in the measurement units of the true intake. The coefficients α and β can therefore also be referred to as scaling factors. Also the random errors ε_i undergo rescaling, to yield standardized random errors ε_i^*. The correlation between the measurements X and true intake T depends on the variance of the standardized random errors relative to the variance of the true intake being measured, as : $\rho_{XT} = 1 / \sqrt{(1 + \sigma^2_\varepsilon/(\beta^2 \sigma^2_T))}$.

Estimation of the error parameters

Suppose one seeks to measure habitual, long-term dietary intake of individuals, using a dietary questionnaire, which is related to true intake as :

$$Q = \alpha_Q + \beta_Q T + \varepsilon_Q \qquad [1].$$

An evaluation of the accuracy (i.e., validity plus precision) of the questionnaire measurements implies that the magnitude of the unknown error parameters α_Q, β_Q and $\sigma^2_{\varepsilon Q}$ should be estimated. Ideally such estimates would be obtained by comparing the individuals' questionnaire assessments with the

corresponding true intake values. However, the true intake values will never be known, but must rather be seen as values of a latent variable (14). It is thus only possible to compare the questionaire assessments with "reference" measurements R, which are related to the same latent variable but which may also contain some error :

$$R = \alpha_R + \beta_R T + \varepsilon_R \quad [2].$$

Throughout this paper, we shall assume that the latent, true intake variable T has a normal distribution, with mean μ_T and variance σ_T^2, and that also all random measurement errors (ε) are normally distributed (with mean 0 and variance σ_ε^2, as defined earlier in the section on the measurement error model). It follows that also the observed measurements Q and R will have a normal distribution. Since the expected measurements $E(Q|T)$ and $E(R|T)$ are both assumed be linearly related to the same latent variable T, under the assumption of joint multi-variate normality of the R,Q,T distribution there must also be a linear relation between the two types of expected measurement:

$$E(Q|T) = \alpha'_Q + \beta'_Q \; E(R|T) \quad [3],$$

where $\alpha'_Q = \alpha_Q - \beta_Q\alpha_R/\beta_R$, and $\beta'_Q = \beta_Q/\beta_R$. For the remainder of this paper it will be assumed that reference measurements can be found without any constant or proportional bias (i.e., α_R=0, and β_R=1.0), so that $\alpha'_Q = \alpha_Q$, and $\beta'_Q = \beta_Q$.

The combination of equations [1], [2] and [3] describes the so called "structural" relation (15,16) that one theoretically expects to observe between the measurements of Q and R. The equations are therefore said to define a structural equations model. In the following sections we shall discuss the possibility of estimating the unknown parameter values in the structural equations model, depending on the number and type of additional measurements available. The parameter estimates can in principle be computed by finding optimal correspondence between the theoretical, multivariate normal distribution of Q,R,X as predicted by the structural equations model (where X stands for any additional measurement involved in the comparison) and the distribution of measurements actually observed. Given our assumption that true and measured intake values are normally distributed, the theoretical and observed multivariate distributions are fully characterized by their first and second moments, that is, by their (theoretical or observed) means, variances and covariances. The parameter values can thus be estimated by fitting the theoretical moments predicted by the structural

equations model to the observed moments estimated from actual measurements in a population sample (14-16). Version 6 of the SAS package for statistical analysis (17) provides a programme for the analysis of structural equations models, the "CALIS" procedure, to obtain maximum likelihood estimates of the parameter values and their confidence intervals. In the Appendix, CALIS programmes are given that were used for the computations of the numerical examples 2 and 3 in this text. In this paper we shall focus on situations where, with additional assumptions, the number of error parameters to be estimated is equal to the number of sample moments. For these situations, explicit forms of the maximum likelihood estimators can be obtained, by simply equating the sample means, variances and covariances to their values predicted by the model. These estimators have been more extensively discussed by Barnett (18), by Jaech (19), and by Dunn (20).

Method 1. Comparison with a single reference measurement

The simplest possible comparison is that between the questionnaire assessment and a single reference measurement. As an example of a reference measurement - often used in dietary validity studies - one may think of a mean intake estimate calculated from a series of weighed food records. We shall assume that random errors of the questionnaire assessments and of the reference measurements are uncorrelated. Given this assumption, and given the structural equations model defined by equations [1]-[3], the pairs of R- and Q-measurements can be considered as a sample of observations from a bivariate normal distribution, of which the means, variances and covariance can be expressed in the various error parameters of interest (see Table 1.A). As the minimal set of sufficient statistics we have the sample moments of the observed R,Q-distribution. Equating the sample moments to the predicted moments of the theoretical, bivariate normal distribution yields five estimating equations, which however are expressed in six unknown parameters :

$$(4.a)\quad \hat{\mu}_T = \bar{R}$$

$$(4.b)\quad \hat{\sigma}^2_T = S^2_R - \sigma^2_{\varepsilon R}$$

$$(4.c)\quad \hat{\beta}_Q = S_{Q,R} \,/\, (S^2_R - \sigma^2_{\varepsilon R})$$

$$(4.d)\quad \hat{\alpha}_Q = \bar{Q} - \bar{R}\,[\, S_{Q,R} \,/\, (S^2_R - \sigma^2_{\varepsilon R})\,]$$

$$(4.e) \quad \hat{\sigma}^2_{\varepsilon Q} = S^2_Q - S_{Q,R}\,[S_{Q,R} / (S^2_R - \sigma^2_{\varepsilon R})]$$

Here, $\bar{R}$, $\bar{Q}$, S^2_R and S^2_Q represent the observed sample means and variances of the R- and Q-measurements, respectively, while $S_{Q,R}$ is their sample covariance.

Only μ_T can be uniquely determined from these estimating equations, as the mean of the reference measurements (equation 4.a). By contrast, the estimates of σ^2_T, α_Q, β_Q, and $\sigma^2_{\varepsilon Q}$ (equations 4.b-4.e) depend on the value of $\sigma^2_{\varepsilon R}$, and can thus only be determined if this value is known. However, the variance of the reference measurement, $\sigma^2_{\varepsilon R}$, cannot be estimated from the equations 4.a-4.e unless the value of at least one of the parameters σ^2_T, α_Q, β_Q, and $\sigma^2_{\varepsilon Q}$ can be assumed to be known a priori. Since there are more unknown parameters than estimating equations, more than one set of parameter estimates can be found for which the predicted moments are equal to the sample moments actually observed. Thus, all model parameters cannot be estimated with the given study design, unless additional (external) information is available on the value of at least one parameter. The model is then unidentifiable. Some traditional approaches ignore this identifiability problem, by assuming that the R-measurements equal the individuals' true intake values, that is, $\sigma^2_{\varepsilon R} = 0$. Substituting 0 for $\sigma^2_{\varepsilon R}$ in equations 4.b-4.e, the usual formulas for linear regression analysis follow. For instance β_Q would be estimated as $S_{Q,R}/S^2_R$, which is the slope of Q regressed on R. The estimates of α_Q and $\sigma^2_{\varepsilon Q}$ are in this case equal to the intercept of the regression line, and the variance of the residual errors of the regression analysis.

Example 1

In a Swedish dietary validity study (21), the usual daily intake of vitamin C was assessed by a dietary questionnaire in a population sample of 107 subjects (men and women combined). The questionnaire assessments were compared with reference measurements, based on the average of two three-day weighed food records. Table 1 shows the estimated sample moments for the Q-, and R-measurements (which had been transformed to improve normality of their distributions (22,23)). The second part of Table 1 shows the error parameter estimates obtained from the equations 4.a-4.e, assuming that the reference measurements had a (close to) perfect correlation with true intake values (i.e., assuming that $\sigma^2_{\varepsilon R} = 0$).

Table 1. Predicted and observed moments, and error parameter estimates for Q-, and R-measurements of vitamin C intake.

Predicted Moments

	Covariance Matrix		Means
	R	Q	
R	$\sigma_T^2 + \sigma_{\varepsilon R}^2$		μ_T
Q	$\beta_Q \sigma_T^2$	$\beta_Q^2 \sigma_T^2 + \sigma_{\varepsilon Q}^2$	$\alpha_Q + \beta_Q \mu_T$

Observed Moments*

	Covariance Matrix		Means
	R	Q	
R	$S_R^2 = 2.00$		$\bar{R} = 6.29$
Q	$S_{Q,R} = 1.42$	$S_Q^2 = 3.32$	$\bar{Q} = 9.68$

* Measurements were transformed to improve normality, using "Box-Cox" (23) power transformations (i.e., using $X_t = (X^{\lambda}-1)/\lambda$, where λ_Q=0.3, and λ_R=0.2)

Parameter Estimates

$\hat{\mu}_T = 6.29$ (6.02,6.56) $\quad \hat{\sigma}_T^2 = 2.00$ (0.81,2.11)

$\hat{\sigma}_{\varepsilon R}^2 = 0$

$\hat{\alpha}_Q = 5.20$ (3.88,6.52) $\quad \hat{\beta}_Q = 0.71$ (0.51,0.91) $\quad \hat{\sigma}_{\varepsilon Q}^2 = 2.31$(1.69,2.93)

(between parentheses are 95 percent confidence intervals)

The estimated scaling factors for the questionnaire indicated the presence of a constant bias (α_Q = 5.20), as well as some proportional bias (β_Q = 0.71). The estimated correlation between questionnaire assessments and the "latent" true intake could be computed as $\hat{\rho}_{QT} = 1 / \sqrt{(1 + \hat{\sigma}^2_Q/(\hat{\beta}^2_Q \hat{\sigma}^2_T))} = 1 / \sqrt{(1 + 2.31/(0.71^2\ 2.00)} = 0.55$. Note, however, that in this particular case the correlation could also have been estimated directly as that between the Q and the R-measurements, ρ_{QR}, since it was assumed that $\sigma^2_{\varepsilon R} = 0$.

The assumption of a perfect correlation between the reference measurements and the underlying true intake is rather a strong one, and the validity of this assumption may be quite doubtful. If in reality $\sigma^2_{\varepsilon R} \neq 0$, the failure to take account of random error in the R-measurements had biased our estimates of each of the error parameters. For instance, the estimate of β_Q was then biased by a factor $\sigma^2_T/(\sigma^2_T+\sigma^2_{\varepsilon R})$, which is known as the "attenuation" bias (6,9). Likewise, α_Q and σ^2_Q were then over-estimated. Therefore, rather than relying on assumptions about the magnitude of the variance $\sigma^2_{\varepsilon R}$, further analyses were performed in order to try to estimate this error variance from additional information.

Method 2. Comparison with repeat reference measurements.

In dietary validity studies it has become common practice to use as a reference measurement an average intake estimate computed from a series of repeat recordings of daily intake (i.e., repeated weighed food records or 24-hour recalls). Therefore, rather than simply assuming that $\sigma^2_{\varepsilon R} = 0$, one might also try to solve the identifiability problem by considering repeat daily recordings as separate measurements. In the validity study on vitamin C intake, for instance, the reference measurement could be considered as an average of two measurements, R_1 and R_2, each based on a three-day food record. Predicted moments of the Q,R_1,R_2 distribution are given in Table 2.A, still assuming that for both R-measurements random errors are independent of those of Q. By equating these to the observed sample moments, six equations expressed in seven unknown parameters are obtained:

(5.a) $\hat{\mu}_T = \frac{1}{2}(\bar{R}_1+\bar{R}_2)$

(5.b) $\hat{\sigma}^2_T = S_{R_1,R_2} - \sigma^2_{\delta R}$

(5.c) $\hat{\beta}_Q = \frac{1}{2}(S_{Q,R_1} + S_{Q,R_2}) \;/\; (S_{R_1,R_2} - \sigma^2_{\delta R})$

(5.d) $\hat{\alpha}_Q = \bar{Q} - [\, \frac{1}{2}(S_{Q,R_1} + S_{Q,R_2}) \;/\; (S_{R_1,R_2} - \sigma^2_{\delta R}) \,]\;\; \frac{1}{2}(\bar{R}_1+\bar{R}_2)$

(5.e) $\hat{\sigma}^2_{\varepsilon Q} = S^2_Q - [\, \frac{1}{2}(S_{Q,R_1} + S_{Q,R_2}) \,]^2 \;/\; (S_{R_1,R_2} - \sigma^2_{\delta R})$

(5.f) $\hat{\sigma}^2_{\varepsilon R} = \frac{1}{2}(S^2_{R_1} + S^2_{R_2}) - S_{R_1,R_2} + \sigma^2_{\delta R}$

Rather than by assuming that the reference measurements do not have any random error at all (i.e., $\sigma^2_{\varepsilon R}=0$), one may now estimate all error parameters on the more relaxed assumption that the random errors of repeat reference measurements are independent (i.e., $\sigma^2_{\delta R} = 0$). Given this additional assumption, we can write : $\sigma^2_{\varepsilon R} = \sigma^2_{\delta R} + \sigma^2_{\gamma R} = \sigma^2_{\gamma R}$. The total random error variance $\sigma^2_{\varepsilon R}$ can thus be estimated from the repeat measurements R_1 and R_2, as the variance of the within-subject random error (equation 5.f). The coefficient β_Q can now be determined as $\hat{\beta}_Q = S_{Q,R}/(S^2_Q - \hat{\sigma}^2_{\gamma R})$. This is identical to the slope of Q regressed on R, with correction for the attenuation bias, $S^2_R/(S^2_R-\sigma^2_{\gamma R})$, due to within-subject random error. Estimates of α_Q and $\sigma^2_{\varepsilon Q}$ are then identical to the intercept and to the variance of residual errors of this corrected regression line.

One could argue that, in spite of the different nature of the Q- and the R_1-measurements, their random errors might not be entirely independent if both measurements are taken very close in time. Freedman et al refined the estimation procedure with repeat reference measurements described above, to take into account a possible covariance between the random errors of the Q- and R_1-measurements (10). The errors of R_2- and Q-measurements were still assumed to be independent, however, as these measurements were taken at a greater distance in time. The predicted moments are given in Table 2.B. Equating these to the observed sample moments yields the following estimating equations:

Table 2.A. Predicted and observed moments, and error parameter estimates, for Q-, and R-measurements (R_1, and R_2) of vitamin C intake.

Predicted Moments

	Covariance Matrix			Means
	R_1	R_2	Q	
R_1	$\sigma_T^2 + \sigma_{\varepsilon R}^2$			μ_T
R_2	σ_T^2 (+ $\sigma_{\delta R}^2$)	$\sigma_T^2 + \sigma_{\varepsilon R}^2$		μ_T
Q	$\beta_Q \sigma_T^2$	$\beta_Q \sigma_T^2$	$\beta_Q^2 \sigma_T^2 + \sigma_{\varepsilon Q}^2$	$\alpha_Q + \beta_Q \mu_T$

Observed Moments *

	Covariance Matrix			Means
	R_1	R_2	Q	
R_1	$S_{R_1}^2 = 2.53$			$\bar{R}_1 = 6.15$
R_2	$S_{R_1,R_2} = 1.47$	$S_{R_2}^2 = 2.55$		$\bar{R}_2 = 6.25$
Q	$S_{Q,R_1} = 1.36$	$S_{Q,R_2} = 1.50$	$S_Q^2 = 3.32$	$\bar{Q} = 9.68$

* Measurements were transformed to improve normality, using "Box-Cox" (23) power transformations (i.e., using $X_t = (X^{\lambda}-1)/\lambda$, where $\lambda_Q = 0.3$, and $\lambda_{R_1} = \lambda_{R_2} = 0.2$).

Parameter Estimates

$\hat{\mu}_T = 6.20$ (5.93,6.46) $\quad \hat{\sigma}_T^2 = 1.47$ (0.91,2.03)

$\hat{\sigma}_{\varepsilon R}^2 = 1.07$ (0.79,1.37)

$\hat{\alpha}_Q = 3.64$ (1.70,5.58) $\quad \hat{\beta}_Q = 0.97$ (0.66,1.28) $\quad \hat{\sigma}_{\varepsilon Q}^2 = 1.93$ (1.26,2.60)

(between parentheses are 95 percent confidence intervals)

(6.a) $\hat{\mu}_T = \frac{1}{2}(\bar{R}_1+\bar{R}_2)$

(6.b) $\hat{\sigma}^2_T = S_{R_1,R_2} - \sigma^2_{\delta R}$

(6.c) $\hat{\beta}_Q = S_{Q,R_2} / (S_{R_1,R_2} - \sigma^2_{\delta R})$

(6.d) $\hat{\alpha}_Q = \bar{Q} - [\, S_{Q,R_2} / (S_{R_1,R_2} - \sigma^2_{\delta R}) \,] \; \frac{1}{2}(\bar{R}_1+\bar{R}_2)$

(6.e) $\hat{\sigma}^2_{\varepsilon Q} = S^2_Q - [\, S_{Q,R_2} \,]^2 / (S_{R_1,R_2} - \sigma^2_{\delta R})$

(6.f) $\hat{\sigma}^2_{\varepsilon R} = \frac{1}{2}(S^2_{R_1} + S^2_{R_2}) - S_{R_1,R_2} + \sigma^2_{\delta R}$

(6.g) $\hat{\sigma}_{\varepsilon R_1,\varepsilon Q} = S_{Q,R_1} - S_{Q,R_2}$

Again, these equations can be solved under the additional assumption that the random errors of repeat reference measurements are uncorrelated ($\sigma^2_{\delta R}=0$).

Example 2

The sample moments of the observed Q,R_1,R_2 distribution in the vitamin C data are given in Table 2.A. The second part of Table 2.A shows the estimates for each of the error parameter estimates, according to first model where $\sigma_{\varepsilon R1,\varepsilon Q}$ was assumed to be equal to 0. Comparison of these estimates with those of example 1 shows that in the first analysis parameters were biased by attenuation, due to within-subject random error in the reference measurements. Correction for such bias in this second analysis resulted in lower estimates of the constant scaling bias (e.g., α_Q = 3.64) and of the random error variance ($\sigma^2_{\varepsilon Q}$ = 1.93). Proportional scaling bias no longer appeared to be present (β_Q = 0.98). The correlation between the questionnaire measurements and the latent variable was estimated to be equal to $\rho_{QT} = 1 / \sqrt{(1 + 1.92/(0.97^2\ 1.47)\ )} = 0.65$.

Parameter estimates according to the model of Freedman et al are given in Table 2.B. The covariance between the random errors of R_1 and Q appeared to be very small, and was even slightly negative ($\sigma_{\varepsilon R1,\varepsilon Q}$ = - 0.13). Consequently, estimated parameter values were almost identical to those in Table 2.A.

Table 2.B. Predicted moments*, and error parameter estimates for Q-, and R-measurements (R_1, and R_2) of vitamin C intake, following Freedman's model.

Predicted Moments

	Covariance Matrix			Means
	R_1	R_2	Q	
R_1	$\sigma_T^2 + \sigma_{\varepsilon R_1}^2$			μ_T
R_2	σ_T^2 (+ $\sigma_{\delta R}^2$)	$\sigma_T^2 + \sigma_{\varepsilon R_2}^2$		μ_T
Q	$\beta_Q \sigma_T^2 + \sigma_{\varepsilon R_1, \varepsilon Q}$	$\beta_Q \sigma_T^2$	$\beta_Q^2 \sigma_T^2 + \sigma_{\varepsilon Q}^2$	$\alpha_Q + \beta_Q \mu_T$

Parameter Estimates

$\hat{\mu}_T = 6.20$ (5.93,6.46) $\quad \hat{\sigma}_T^2 = 1.47$ (0.91,2.02)

$\hat{\sigma}_{\varepsilon R}^2 = 1.08$ (0.79,1.37) $\quad \hat{\sigma}_{\varepsilon R_1, \varepsilon Q} = -0.13$ (-0.55,0.29)

$\hat{\alpha}_Q = 3.35$ (1.15,5.55) $\quad \hat{\beta}_Q = 1.02$ (0.66,1.37) $\quad \hat{\sigma}_{\varepsilon Q}^2 = 1.79$ (0.95,2.63)

(between parentheses are 95 percent confidence intervals)

* Observed moments were the same as in Table 2.A.

The assumption that $\sigma^2_{\delta R} = 0$ means that the full set of error parameters can be estimated without bias, but only if the reference measurement is repeated at least once. This is a more relaxed assumption than that in Example 1, where it was necessary to assume that $\sigma^2_{\varepsilon R}=0$; that is, the correlation between measured and true intake values was assumed to be perfect even for a single R-measurement. However, there may also be doubts about the validity of this more relaxed assumption. If in reality $\sigma^2_{\delta R} \neq 0$, all parameter estimates except that of μ_T would be biased. For instance, β_Q is estimated as $\frac{1}{2}(S_{Q,R1} + S_{Q,R2})/S_{R1,R2}$ (equation 5.c). Filling in the predicted moments from Table 2.A, it can be easily seen that the expected value of this β_Q-estimate would be equal to $\beta_Q\sigma^2_T/(\sigma^2_T+\sigma^2_{\delta R})$, where the factor $\sigma^2_T/(\sigma^2_T+\sigma^2_{\delta R})$ expresses a residual attenuation bias.

Method 3. Comparison with a reference measurement plus a third type of measurement

Barnett (18) showed that the problem of identifiability of error parameters can also be solved by comparing the questionnaire assessments with at least two different types of measurement. Instead of taking duplicate reference measurements by a similar method, for instance based on weighed food records or 24-hour recalls, one may also obtain a third measurement using a very different method such as a biochemical marker (M). The assumption that the random errors ε are independent for each pair of measurements is then more likely to be valid. Predicted moments of the Q,R,M -distribution are given in Table 3. Equating these to the observed sample moments yields nine estimating equations, expressed in an equal number of unknown parameters. The parameters can therefore all be estimated without additional assumptions:

(7.a) $\hat{\mu}_T = \bar{R}$

(7.b) $\hat{\sigma}^2_T = (S_{Q,R}\, S_{M,R}) \,/\, S_{Q,M}$

(7.c) $\hat{\alpha}_Q = \bar{Q} - (S_{Q,M}\, \bar{R}) \,/\, S_{M,R}$

(7.d) $\hat{\alpha}_M = \bar{M} - (S_{Q,M}\, \bar{R}) \,/\, S_{Q,R}$

(7.e) $\hat{\beta}_Q = S_{Q,M} \,/\, S_{M,R}$

(7.f) $\hat{\beta}_M = S_{Q,M} \,/\, S_{R,Q}$

$$(7.g)\quad \hat{\sigma}^2_{\varepsilon R} = S^2_R - (S_{Q,R}\, S_{M,R}) \,/\, S_{Q,M}$$

$$(7.h)\quad \hat{\sigma}^2_{\varepsilon Q} = S^2_Q - (S_{Q,M}\, S_{Q,R}) \,/\, S_{M,R}$$

$$(7.i)\quad \hat{\sigma}^2_{\varepsilon M} = S^2_M - (S_{M,R}\, S_{Q,M}) \,/ S_{Q,R}$$

It is worth noting that we can now estimate the variance of the total random error of the R-measurements, $\sigma^2_{\varepsilon R}$, which includes the variance of random biases $\sigma^2_{\delta R}$. This underlines that it is not essential that the reference measurements provide a fully accurate estimate of the individuals' ranking by true intake level. Even when the total random error variance $\sigma^2_{\varepsilon R}$ is relatively large, all error parameters are expected to be estimated without bias. It should be kept in mind, however, that the estimates will then also have relatively wide confidence intervals unless the dietary validity study is based on a very large number of individuals.

The biomarker assessment, M, can be seen as merely an "instrumental" variable (15,16), which makes it possible to estimate the "true" regression of the Q- on the R-measurements with adjustment for all attenuation bias due to the total random error in the reference measurements. In other words, the biomarker assessment allows to estimate the relation between the Q-assessments and the latent variable, expressed in the measurement units of R. It should be noted that it is not necessary to know the quantitative, functional relation between the M-measurements and true intake T. In fact, any third M can be used as an instrumental variable as long as it has a linear relationship with the underlying latent variable, and random errors that are independent of those of R and Q. For many biomarkers of dietary intake, such as the blood concentration of a particular vitamin, or the fatty acid composition of a tissue biopsy, the quantitative relation to absolute intakes is quite unclear. Still, such markers can provide ideal instrumental measurements since they can have good correlations with true intake levels (24-26). Theoretically, one could even envisage the use of body mass index or total energy expenditure as potential instrumental measurements if their relation with the underlying true intake factor is strong enough. However, one would expect a low correlation between the instrumental measurement and the latent variable to result in relatively large confidence intervals, even though parameter estimates would still be expected to be unbiased.

Table 3. Predicted and observed moments, and error parameter estimates for Q-, R-, and M-measurements of vitamin C intake.

Predicted Moments

	Covariance Matrix			Means
	R	Q	M	
R	$\sigma_T^2 + \sigma_{\varepsilon R}^2$			μ_T
Q	$\beta_Q \sigma_T^2$	$\beta_Q^2 \sigma_T^2 + \sigma_{\varepsilon Q}^2$		$\alpha_Q + \beta_Q \mu_T$
M	$\beta_M \sigma_T^2$	$\beta_Q \beta_M \sigma_T^2$	$\beta_M \sigma_T^2 + \sigma_{\varepsilon M}^2$	$\alpha_M + \beta_M \mu_T$

Observed Moments *

	Covariance Matrix			Means
	R	Q	M	
R	$S_R^2 = 2.00$			$\bar{R} = 6.29$
Q	$S_{Q,R} = 1.42$	$S_Q^2 = 3.32$		$\bar{Q} = 9.68$
M	$S_{M,R} = 5.63$	$S_{Q,M} = 5.48$	$S_M^2 = 39.87$	$\bar{M} = 17.70$

* All measurements were transformed to improve normality, using "Box-Cox" (23) power transformations (i.e., using $X_t = (X^{\lambda}-1)/\lambda$, where λ_Q=0.3, λ_R=0.2, and λ_M=0.7).

Parameter Estimates

$\hat{\mu}_T = 6.29$ (6.02,6.56) $\quad \hat{\sigma}_T^2 = 1.46$ (0.81,2.11)

$\hat{\sigma}_{\varepsilon R}^2 = 0.54$ (0.12,0.96)

$\hat{\alpha}_Q = 3.57$ (1.44,5.68) $\quad \hat{\beta}_Q = 0.97$ (0.63,1.31) $\quad \hat{\sigma}_{\varepsilon Q}^2 = 1.92$ (1.30,2.58)

$\hat{\alpha}_M = -6.57$ (-14.49,1.35) $\quad \hat{\beta}_M = 3.86$ (2.61,5.11) $\quad \hat{\sigma}_{\varepsilon M}^2 = 18.2$ (10.5,25.8)

(between parentheses are 95 percent confidence intervals)

Example 3

In the validity study on vitamin C intake, a third estimate of the validity of the questionnaire assessment of vitamin C intake was obtained by comparison with the overall average of R-measurements, and a biochemical marker, M, which was an average of six different measurements of vitamin C concentration in blood serum. The moments of the observed Q,R,M distribution and estimates of the error parameters are given in Table 3.

Interestingly, the estimates from the third analysis were virtually identical to those in Tables 2.A and 2.B, indicating that in fact there was already no residual attenuation bias left using method 2. In this final analysis, the total random error variance for a reference measurement based on a six-day food record was estimated to be equal to $\sigma^2_{\varepsilon R} = 0.54$. This is exactly half of the variance of the within-subject random error for a three-day record, estimated in the previous analysis ($\sigma^2_{\gamma R} = 1.08$; see Table 2.A). We therefore concluded that indeed $\sigma^2_{\delta R} \simeq 0$. This particular data example thus seems to confirm the common assumption that day-to-day variations in the individuals' true intake are virtually the only source of random error, when repeat weighed records are used for the assessment of usual, long-term intake.

Precision of the parameter estimates

Although parameter estimates were almost identical in Tables 2.A and 3 (since apparently $\sigma^2_{\delta R} \simeq 0$), and in spite of the fact that in Example 3 the overall information per subject had been increased by adding the M-measurements, confidence intervals were larger for the estimates obtained by method 3. This can be explained by the fact that using method 3 a larger number of parameters has to be estimated.

According to the estimates of Table 3, the correlation between M and T equalled $1/\sqrt{(1+18.2/(3.86^2 * 1.46)} = 0.74$. The question arose whether, if this correlation had been even stronger, method 3 could have given more efficient estimates than method 2. This was investigated by a simple simulation, modifying the value of the sample variance of M-measurements into 22.73 instead of the value of 39.87 that was truly observed. This modification left all parameter estimates of Table 3 unaffected, except that $\sigma^2_{\varepsilon M}$ was now estimated to be equal to zero, indicating a perfect correlation between M and T. In addition, confidence intervals of various parameter estimates had become narrower, and were also slightly narrower than for the estimates

obtained by method 2. For instance, β_Q was estimated to be equal to 0.97±0.25, against 0.97±0.31 in Table 2.

In all examples shown, the confidence intervals were rather large, indicating that validation studies should be based on a larger number of observations. For the parameter estimates obtained from equations 7.a-7.i, closed form formulas for the estimation of confidence intervals have been given by Barnett (18). Such formulas indicate how random errors in each of the measurements affect the precision of the parameter estimates, and may thus be helpful in determining sample size requirements for dietary validity studies, in terms of the total number of individuals to be included, or of numbers of repeat assessments to be taken on each individual. Discussion of such design options is, however, beyond the scope of this paper.

Discussion

We have discussed the estimation of the magnitude of measurement error in assessments of the habitual dietary intake of individuals in terms of structural equation models, which belong to the more general class of latent variable models (14,20). We have shown that, to estimate accurately the scale of error in dietary questionnaire assessments, these must be compared with at least two additional measurements. The first - which we constantly refer to as the "reference" measurement - is by definition unbiased at a group level, conditional on the true intake value (i.e., α_R=0 and β_R=1.0). The second additional measurement should at least provide another independent estimate of the individuals' ranking by intake level, and can either be based on a repeated reference measurement (assuming that the errors of repeat reference measurements are uncorrelated) or on a third method such as a biochemical marker.

A major requirement for validity studies is that a comparison be made between measurements that do not tend to have similar errors for the same individuals (i.e., the errors are independent). Any correlation between these measurements will then be due only to the fact that they relate to the same latent variable. If errors cannot be assumed to be independent, it will no longer be clear whether a correlation between measurements exists because they each really measure the same thing, or merely because the errors are correlated. Mathematically, this will be expressed as a problem of identifiability: if it is suspected that errors are correlated, the

structural equations model should also incorporate parameters for the covariances between errors. As a consequence, there will then be more parameters to be estimated than there are sample moments. Even if it is possible to test for the independence of the errors of some measurements, such a test will always have to rely on the assumption of independent errors for other measurements that were taken for the same individuals. For instance, parameters in the model of Freedman et al. would become unidentifiable if one does not assume that the errors of the Q-measurements are independent of those of at least one of the two R-measurements. A validity study must therefore always rely on the assumption that, for at least a certain number of measurements, errors are uncorrelated. The decision on which methods can be considered different enough to have independent errors will depend on the researcher's understanding of the nature of each method used, and their potential sources of error. In this paper we assumed that there may be three main categories of methods for which random errors will be independent, namely: 1) dietary questionnaires to obtain an immediate estimate of usual food consumption, 2) methods based on the recording of actual food consumption on one or more randomly selected days, and 3) biochemical markers of dietary intake.

The procedures discussed in this paper to estimate error parameters and their confidence intervals are valid on the assumption that the distributions of true, as well as measured, intake are normal, and that associations between different measurements are linear. In practice, distributions of dietary intake measurements are often positively skewed, approaching log-normality (23,27). Scatterplots often show a widening pattern with increasing values of intake measurements, suggesting that such log-normality may be due to a larger variance of random error in the larger values of each type of measurement. We used "Box-Cox" power transformations to correct for the positive skewness of dietary intake measurements (22,23). It is believed that, while the overall distributions of measurements will be normalized as a result of such transformations, at the same time the random errors will become more homoscedastic. It may remain unclear however, how transformations separately influence the distributions of the latent variable and of the random errors, and how they might affect the assumption of linearity of relations between different types of measurement. Since random error is generally a predominant source of variation, scatterplots

can provide only limited information about the shape of the structural relation between measurements.

The various parameter estimates given in this paper can be useful during the planning phase of epidemiological studies on diet in relation to disease. Validity studies will allow the evaluation and selection of an optimal dietary questionnaire for the assessment of dietary exposures of interest, before it is applied in a large-scale study. The questionnaire's capacity to rank individuals by true intake level is described by the ratio $\sigma^2_{\varepsilon Q}/\sigma^2_T$, while estimates of α_Q and β_Q will indicate whether the questionnaire measurements contain constant or proportional scaling bias. In addition, knowledge of the variance of the true intake distribution, σ^2_T, may be an essential element in the determination of sample size requirements for cohort studies (28).

At the stage of analysis of the epidemiological study results, the estimated error model parameters could in principle also be used to adjust point and interval estimates of relative risk for biases due to error in dietary questionnaire assessments of exposure (29,30). A correction factor $(\sigma^2_T+\sigma^2_{\varepsilon Q})/\sigma^2_T$ would be needed to adjust a logistic regression coefficient for attenuation bias, while a multiplication by β_Q would adjust for over- or under-estimation of relative risk due to the proportional scaling bias in the questionnaire measurements. However, in order to make these adjustments it is not really necessary to first obtain estimates of β_Q, $\sigma^2_{\varepsilon Q}$, and σ^2_T separately. Rosner et al (30) have described how the overall correction factor $\beta_Q(\sigma^2_T+\sigma^2_{\varepsilon Q})/\sigma^2_T$ can also be estimated as the reciprocal of the slope of a linear regression of a single reference measurement R on the dietary exposure assessments (assuming that $E(R|T) = T$). A validity study, based on more than one additional intake measurement, is not needed therefore for the sole purpose of adjusting relative risk estimates. In fact, it is then the measurement of disease status which plays the role of a third measurement related to the latent intake variable.

We conclude that, even in the absence of truly valid reference measurements, it is possible to evaluate the validity of dietary intake assessments using latent variable models. However, for sufficient precision, validity studies should be based on larger numbers of observations than has been usually the case, while the robustness of the estimation procedures

relying on assumptions of normality may also need further evaluation.

APPENDIX: SAS Calis programs used.

```
/* Data input from Table 2.a */
data vitc(type = cov);
  input _type_ $ _name_ $ R1 R2 Q;
  cards;
n . 107 107 107
mean . 6.15 6.25 9.68
cov R1 2.53 . .
cov R2 1.47 2.55 .
cov Q 1.36 1.50 3.32 ;

/* Computations for Table 2.a */
proc calis ucov aug stderr data = citc;
  lineqs R1 = meanR intercep + fT + eR1,
         R2 = meanR intercep + fT + eR2,
         Q  = meanR intercep + betaQ fT + eQ;
  parameters alphaQ;
       meanQ = alphaQ + betaQ*meanR;
  std eR1 = veR,
      eR2 = veR,
      eQ  = veQ,
      fT  = vT;
run;

/* Computations for Table 2.b */
proc calis ucov aug stderr data = citc;
  lineqs R1 = meanR intercep + fT + eR1,
         R2 = meanR intercep + fT + eR2,
         Q  = meanR intercep + betaQ fT + eQ;
  parameters alphaQ;
       meanQ = alphaQ + betaQ*meanR;
  std eR1 = veR1,
      eR2 = veR2,
      eQ  = veQ,
      fT  = vT;
  cov eR1 eQ = cov:;
run;


/* Data input from Table 3 */
data vitc(type = cov);
  input _type_ $ _name_ $ R Q M;
  cards;
n . 107 107 107
mean . 6.29 9.68 17.70
cov R1 2.00 . .
cov R2 1.42 3.32 .
cov Q  5.63 5.48 39.87
;
```

```
/* Computations for Table 3 */
proc calis ucov aug stderr data = citc;
  lineqs R = meanR intercep + fT + eR,
         Q = meanQ intercep + betaQ fT + eQ,
         M = meanM intercep + betaM fT + eM;
  parameters alphaQ alphaM;
       meanQ = alphaQ + betaQ*meanR;
       meanM = alphaM + betaM*meanR;
  std eR = veR,
      eQ = veQ,
      eM  = veM,
      fT  = vT;
run;
```

References

1. Block G. A review of validations of dietary assessment methods. Am J Epidemiol 1982;115:492-505.
2. Cameron ME, van Staveren WA (eds.). Manual on methodology for food consumption studies. Oxford: Oxford University Press, 1988.
3. Burema J, van Staveren WA. Validation of the dietary history method. In : Kohlmeier L (ed.). The diet history method. Proceedings of the 2nd Berlin meeting on nutritional epidemiology. London: Smith-Gordon, 1991.
4. Lee-Han H, McGuire V, Boyd NF. A review of the methods used by studies of dietary measurement. J Clin Epidemiol 1989;42:269-79.
5. Balogh M, Kahn HA, Medalie JH. Random repeat 24-hour dietary recalls. Am J Clin Nutr 1971;24:304-10.
6. Beaton GH, Milner J, Corey P, McGuire V, et al. Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Am J Clin Nutr 1979;32:2546-59.
7. Sempos CT, Johnson NE, Smith EL, Gilligan C. Effects of intraindividual and interindividual variation in repeated dietary records. Am J Epidemiol 1985;121:120-30
8. Nelson M, Black AE, Morris JA, Cole T. Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. Am J Clin Nutr 1989;50:155-67.
9. Rosner B, Willett WC. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. Am J Epidemiol 1988;127:377-88.
10. Freedman LS, Carroll RJ, Wax Y. Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. Am J Epidemiol 1991;134:310-20.
11. Livingstone MBE, Prentice AM, Strain JJ, Coward WA, Black AE, Barker ME, McKenna PG, Whitehead RG. Accuracy of weighed dietary records in studies of diet and health. Br Med J 1990;300:708-12.
12. Bandini L, Schoeller DA, Cyr H, Dietz WH. Validity of reported energy intake in obese and non-obese adolescents. Am J Clin Nutr 1990;52:421-25.
13. Willett W. An overview of issues related to the correction of non-differential exposure measurement error in epidemiological studies. Stat Med 1989;8:1031-40.
14. Everitt BS. An introduction to latent variable models. London: Chapman and Hall, 1984.

15. Kendall MG, Stuart A. Functional and structural relationship. Chapter 29 in: The advanced theory of statistics, Volume 2. London: Griffin & Company, 1967:375-418.
16. Madanski A. The fitting of straight lines when both variables are subject to error. J Am Stat Assoc 1959;54:173-205.
17. SAS Institute Inc. SAS/STAT User's Guide, Version 6, Volume 1, 1989.
18. Barnett VD. Simultaneous pairwise linear structural relationships. Biometrics 1969;25:129-42.
19. Jaech JL. Statistical analysis of measurement errors. Exxon Monograph. New York: John Wiley & Sons, 1985.
20. Dunn G. Design and analysis of reliability studies. The statistical evaluation of measurement errors. New York: Oxford University Press, 1989.
21. Callmer E, Riboli E, Lindgarde F, Saracci R. Malmö methodological study on dietary assessment methods. Study design and dietary habits (unpublished).
22. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc B 1964;26:211-252.
23. Borrelli R, Cole TJ, Di Biase G, Contaldo F. Some statistical considerations on dietary assessment methods. Eur J Clin Nutr 1989;43:453-463.
24. Riboli E, Rönnholm H, Saracci R. Biological markers of diet. Cancer Surveys 1987;6:685-718.
25. Hunter D. Biochemical indicators of dietary intake. Guest chapter in: Willett W. Nutritional epidemiology. New York: Oxford University Press, 1990:375-418.
26. Kok F, van 't Veer P, eds. Biomarkers of dietary exposure. Proceedings of the 3rd meeting on nutritional epidemiology. London: Smith-Gordon, 1991.
27. Emrich LJ, Dennison D, Dennison K. Distributional shape of nutrition data. J Am Diet Assoc 1989;89:665-670.
28. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. Am J Epidemiol 1990;132:1185-95.
29. Armstrong B, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. Stat Med 1989;8:1151-63.
30. Rosner BA, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;8:1051-69.

Chapter 3

"Validation" and "calibration" of dietary intake assessments in prospective cohort studies.

This chapter has been submitted for publication, by the authors R. Kaaks, E. Riboli, and W.A. van Staveren.

Abstract

To evaluate the accuracy of dietary intake measurements in prospective cohort studies on diet, it has been proposed that sub-studies be conducted in order to: 1) correct relative risk estimates for biases due measurement error, and 2) account for statistical power losses when estimating the sample size requirements of the cohort. Usually, the sub-study takes the form of a "validity" study, based on a small group of volunteers, using multiple days of food intake records per subject as reference measurements. In this paper it is shown that, when relative risks are estimated for scaled, absolute intake differences rather than for quantile categories, a "calibration" study based on only a single day's food intake record (but on a larger number of subjects), can provide sufficient information as a reference measurement. A major advantage of calibration studies is that they can be conducted more easily on a representative sample of cohort participants. In addition, it is shown that, for a given number of daily intake records collected, a calibration study is statistically most efficient when it includes a maximum number of subjects, with only a single intake record each.

Introduction

A major limitation of epidemiologic research on diet in relation to chronic diseases such as cancer is the difficulty of obtaining accurate measurements of individuals' habitual, long-term intake levels of foods and nutrients. Correlations between dietary questionnaire measurements and true, long-term intake values are generally estimated to be lower than 0.7 (1,2), which implies that at least half of the variation in intake measurements is due to random errors. As a result of these relatively low correlations, relative risks indicating a relation between dietary intake patterns and the occurrence of disease tend to be underestimated, and the probability of observing a real, statistically significant relation is reduced (3,4). To estimate the magnitude of these effects, it has been recommended that epidemiologic investigations, and in particular prospective cohort studies, should incorporate sub-studies to evaluate the accuracy of dietary questionnaire measurements (5).

In the past, sub-studies on the accuracy of dietary questionnaire measurements have mostly taken the form of validity studies, based on a small group of 100 to 200 volunteers, using daily food intake records as reference measurements (1,2). Repeated records are taken for each subject to improve the precision of the validity study, and to estimate the correlation between the questionnaire measurements and the individuals' true habitual intake values (with adjustment for attenuation biases due to random errors in the reference measurements) (6,7). This correlation coefficient is then taken as the main criterion to evaluate, before starting the main epidemiologic study, whether a newly developed questionnaire instrument measures habitual diet sufficiently accurately or whether sections of it should be improved for specific food groups or nutrients. If more than one type of questionnaire is tested, a validity study can be used to select the version which yields the most accurate measurements (2,3). The estimated correlation coefficient is often used also for the subsequent planning and analysis of the main cohort study in order to:

a) account for inaccuracy of dietary intake measurements when estimating the statistical power or sample size requirements for the main epidemiologic study (3,8); and
b) adjust for attenuation biases in relative risk estimates, which are usually expressed for quantiles (mostly quartiles, or quintiles) of the intake distribution (3,4).

A problem with dietary validity studies is that keeping dietary intake records for many days is a considerable burden. Probably only subjects who are particularly interested in diet and who are motivated to respond accurately to the questionnaires will agree to take part. It therefore seems likely that the correlation between questionnaire measurements and true intake values in the validity study is stronger than in the overall study population. There is consequently a risk of overestimating the statistical power of the main epidemiologic study and underestimating the attenuation biases in relative risk estimates. This hypothesis is supported by recent results from the New York University Women's Health Study (9) where the reproducibility of dietary questionnaire measurements was better among participants who later agreed to fill out the questionnaire for a third time, than among those who did not (10).

In the present paper we review the requirements for sub-studies on the accuracy of dietary questionnaire measurements, focusing on the use of such studies for objectives a) and b) above. It will be shown that, if relative risks are estimated for scaled, quantitative differences in intake level rather than for quantiles of the population distribution of intake levels, both objectives can be met by a calibration study, based on the "linear approximation" method described by Rosner et al. (11). The advantage of this approach is that such calibration studies require only a single day's intake record per subject as a reference measurement, and may therefore be conducted more easily on a representative sub-sample of the study population. A theoretical example is given to illustrate that a calibration sub-study can be conducted ideally during the initial ("pilot") phase of a prospective cohort study.

Effects of dietary measurement errors: bias and statistical power

Bias: Suppose that within a given study cohort the relation between the incidence rate of a given disease (e.g., a specific form of cancer), and the habitual intake T of a dietary factor (e.g. a particular nutrient) is given by:

$$\log(\text{disease rate at intake level } T) = \text{constant} + \theta T \qquad [1].$$

In this exponential risk model, commonly adopted in the analysis of individual-based epidemiologic studies, the slope parameter θ is the logarithm of the relative risk of disease for a unit difference in intake T. The slope parameter can be estimated for instance by logistic regression,

using a case-control study nested within the cohort (12), or by a Poisson type of regression if the analysis is based on the total number of person years observed in the cohort (13). We shall assume that the relation between questionnaire measurements of the intake level of a given nutrient and the individuals' true habitual intake levels is correctly described by a linear measurement error model:

$$Q = \alpha_Q + \beta_Q T + \varepsilon_Q \qquad [2].$$

In this model, which is discussed in more detail elsewhere (7), the coefficients α_Q and β_Q represent constant and proportional scaling biases, which occur, respectively, if individuals tend to over- or underestimate intake by some constant amount, or by an amount proportional to the true intake value itself. The term ε_Q indicates a random error which, at a group level, and conditional on the true intake level T being measured, has zero mean and variance $\sigma^2_{\varepsilon Q}$. Using questionnaire measurements Q instead of true intake values T, the expected log relative risk estimate will be equal to

$$E[\hat{\theta}^*] = \lambda\theta \ , \qquad \text{where } \lambda = \rho^2_{QT}/\beta_Q \qquad [3].$$

This estimate will thus be biased by a factor λ, which is the inverse of the proportional scaling factor β_Q, multiplied by an attenuation factor ρ^2_{QT} which is the square of the correlation between measured and true intake values.

Statistical power: In conjunction with the attenuation of relative risk estimates, an imperfect correlation ρ_{QT} leads to a loss of statistical power for a test of association between the dietary intake factor and disease risk. If we assume that the population distribution of the true intake values is Normal – with mean μ_T and variance σ^2_T – and that the rate of disease incidence is low (as is the case for most specific types of cancer), the power of a test of the null hypothesis H_0:θ=0 is approximately equal to that of a t-test on a mean intake difference between cases and controls (14,15). Thus, for a case-control study nested within the cohort, in which dietary questionnaire measurements of an expected number of D cases are compared to those of j times as many controls, the statistical power can be derived as

$$\text{power} = \Phi\left(\frac{\theta\rho_{QT}\sigma_T}{\sqrt{(j+1)/j}} \sqrt{D} - Z_{\alpha/2} \right) \quad [4].$$

Here $Z_{\alpha/2}$ denotes the 100(1-α/2) centile of the standard normal distribution, and Φ(U) is the probability that a standard normal variate is smaller than U (16). Likewise, the number of cases required to reach a minimal statistical power 1-β (where β here denotes the probability of a type II error) can be derived as:

$$D = (j+1)/j \left(\frac{Z_{\alpha/2}+Z_{\beta}}{\theta\rho_{QT}\sigma_T} \right)^2 \quad [5].$$

Note that in a full cohort analysis there will be many more controls than cases, so that the factor (j+1)/j approaches 1.0.

In reality, the size of a newly planned cohort study can be determined to a large degree by pragmatic considerations. For example, it may be possible to conduct the cohort study within an existing programme originally designed for a different purpose, such as breast cancer screening or blood donation (17), but which provides an economical infrastructure for data collection and follow-up. In this type of situation, where the expected number of cases is determined beforehand by the given cohort size and the planned duration of follow-up, equation 4 can be used to evaluate whether the power to detect a diet-disease association will be high enough to make the study worthwhile. Alternatively, there may be situations where the size of the cohort can be extended to increase the study power. Equation 5 can then be used to calculate the number of cases required to reach a minimum study power. In either case, we must specify a realistic value for the product $\theta\rho_{QT}\sigma_T$, which is composed of three unknown parameters: θ, the logarithm of the relative risk for one unit difference in intake level; σ_T, the standard deviation of the true intake distribution; and ρ_{QT}, the correlation between questionnaire measurements and true intake levels. This product measures the strength of the association between disease risk and the questionnaire measurements of intake, as is illustrated by the fact that it relates to an expected odds ratio of disease between given quantiles of the measured intake distribution (see appendix in Kaaks et al. (18)).

Given our primary interest in obtaining an unbiased estimate of the log

relative risk θ, a natural approach to specifying the product $\theta\rho_{QT}\sigma_T$ is to define first a minimum value for the log relative risk, θ_R, that would be considered of etiologic or public health relevance. This θ_R-value can be defined, for instance, for a given absolute intake difference that is known to exist between subgroups of a given population (15). Alternatively, the θ_R-value may be based on an a priori estimate, obtained for instance from international correlation studies, as has been illustrated for studies on fat intake in relation to cancers of the colon or breast (15,19). Having defined the θ_R-value, we need an estimate of the magnitude of the remaining product $\rho_{QT}\sigma_T$. This smaller product represents the amount of between-person variation in true intake levels that is accurately distinguished – or "predicted" – by the dietary questionnaire measurements; that is, $(\rho_{QT}\sigma_T)^2 = \mathrm{Var}(E[T|Q])$.

In the following paragraphs we shall discuss how to estimate not only the bias factor λ but also the predicted intake variance $(\rho_{QT}\sigma_T)^2$ using the information from additional, unbiased reference measurements, obtained in a sub-sample of cohort participants.

CALIBRATION STUDIES: estimating the bias factor, and the predicted variation in true intake level

Intuitively it can be easily seen that the bias factor λ in equation 3 should be similar to the slope of true intake values T regressed on the questionnaire measurements Q. This led Rosner et al. (11) to describe the following method, known as "linear approximation", to adjust for this bias in log relative risk estimates. In a representative sub-sample of cohort participants, additional "reference" measurements are taken which, at a group level, can be assumed to be free of scaling bias (i.e., $\alpha_R=0$, and $\beta_R=1.0$), and whose errors can be assumed to be independent of those of the baseline questionnaire measurements. The linear approximation method then consists of: 1) estimation of a crude log relative risk estimate $\hat{\theta}^*$, for instance by logistic regression; 2) estimation of the bias factor λ as the slope of a normal linear regression of reference measurements R on questionnaire measurements Q; and 3) estimation of a corrected log relative risk estimate as $\hat{\theta}=\hat{\theta}^*/\hat{\lambda}$. This approach can also be seen as a linear rescaling of the questionnaire measurements by a factor λ (that is, using transformed measurements $Q'=\hat{\lambda}Q$), so that an unbiased estimate of the log

relative risk is obtained by regression of disease status (as a binary outcome variable) on the rescaled measurements. We shall therefore refer to this approach as a "calibration" procedure, and the bias factor λ as a "calibration factor".

The variance of the "calibrated" questionnaire measurements will be equal to $Var(Q') = \lambda^2 Var(Q) = (\rho_{QT}\sigma_T)^2$ (Appendix A), and thus provides an estimate of the amount of between-subject variation in true intake level that is predicted by the baseline questionnaire measurements (i.e., $Var(Q')$ is an estimate of $Var(E[T|Q])$). On the basis of this predicted variation, an estimate can be obtained of the log relative risk θ, and the null hypothesis of "no diet-disease association" ($H_0:\theta=0$) can be tested. Thus, if the calibration study is conducted early in the recruitment phase of a prospective cohort study (e.g., as part of an overall feasibility study), the estimate of the predicted intake variance can be used to calculate the sample size requirements of the cohort.

The reference measurements may be obtained for instance by means of weighed food records, or 24-hour recalls. It is important to note that random errors in the reference measurements (ε_R) are not expected to cause bias in the λ-estimate (attenuation bias depends only on random error in the predictor variable, that is, in the questionnaire measurements). Reference measurements can therefore be based on only a single day's intake record per subject, even though this may be relatively unreliable as a measurement of an individual's long-term, habitual intake level. For the calibration to be precise, however, a sufficient total number of daily intake records should be obtained in the entire sub-study, either by increasing the number of study participants, or by taking repeat records for each participant. In dietary validity studies it is common practice to obtain multiple daily intake records per individual. For a calibration study conducted on a sub-sample of participants in a cohort study, however, it can be shown that, for a given total number of intake records collected, estimates of λ and of the predicted variance $Var(E[T|Q])$ will be most precise when the calibration study is based on a maximum number of individuals, with only a single intake record each (Appendix B). For example, instead of collecting 14 days of weighed food records from 100 subjects, a calibration study will be statistically more efficient if it includes 1400 subjects with only a single record each.

VALIDITY STUDIES: estimating the correlation between questionnaire measurements and true intake values

Although a single reference measurement per individual (e.g., based on a single food record, or a single 24-hour recall) can provide sufficient information to estimate the calibration factor λ as well as the predicted intake variance $Var(E[T|Q])$, this information will not be sufficient for separate estimation of the parameters ρ^2_{QT}, β_Q, and σ^2_T. It will thus be impossible to determine whether bias in relative risk estimates is mainly the result of random dietary measurement errors (i.e., a low correlation ρ_{QT}), or of proportional scaling bias as well. Likewise, it will not be possible to evaluate whether a small predicted intake variance is mainly the result of a low correlation between questionnaire measurements and true intake values, or whether it reflects a small between-subject variation in true intake level. In other words, the loss of statistical power due to random measurement error cannot be estimated.

Studies that do allow a separate estimation of the parameters ρ^2_{QT}, β_Q, and σ^2_T can be referred to as "validity" studies, as they permit a distinction between variation in the questionnaire measurements that reflects true intake differences, and variation which is due to random errors. As discussed in detail elsewhere (7), a major requirement for validity studies is that the questionnaire measurements must be compared with a minimum of two additional intake measurements, at least one of which should be a reference measurement (R) without scaling bias (i.e., $R=T+\varepsilon_R$), whereas the other can be either a repeat reference measurement (R_2), or a biochemical marker of intake. It is also vital that all three measurements have mutually independent random errors. In the past, dietary validity studies have most often been based on a comparison with k repeat reference measurements, R_i (i=1,....,k), each obtained by means of a weighed food record or a 24-hour diet recall. Then, assuming a zero covariance between the errors of repeat reference measurements, we can obtain the following estimates:

$$\hat{\sigma}_T^2 = Cov(R_i,R_j) \qquad [6.a],$$

$$\hat{\rho}_{QT} = Cov(Q,\bar{R})/\sqrt{[Var(Q)Cov(R_i,R_j)]} \qquad [6.b],$$

$$\hat{\beta}_Q = \mathrm{Cov}(Q,\bar{R})/\overline{\mathrm{Cov}}(R_i,R_j) \qquad [6.c],$$

where $\bar{R} = \sum_{i=1}^{k} R_i$ is the average of the k repeat reference measurements of each given individual, and $\overline{\mathrm{Cov}}(R_i,R_j)$ is the mean covariance between repeat reference measurements.

The estimate of the true intake variance σ_T^2 (equation 6.a) is equivalent to the estimated between-subject variance of reference measurements as usually obtained in an analysis of variance using individuals as a grouping factor. Likewise, the estimates of the correlation ρ_{QT} (equation 6.b), and of the proportional scaling bias β_Q (equation 6.c), are equivalent to those which can be obtained by linear correlation and regression analysis with adjustment for attenuation biases due to random error in the reference measurements (2).

Once separate estimates of the correlation ρ_{QT} and of the proportional scaling bias β_Q are obtained from a validity study, they can be used to adjust crude log relative risk estimates for, respectively, attenuation bias and scaling bias. However, the combination of these two corrections is equivalent to a "calibration" adjustment, where the calibration factor is estimated as the slope of the individuals' mean reference measurements ($\bar{R}$) regressed on questionnaire measurements. This can be easily seen from equations [6.b] and [6.c], since:

$\hat{\rho}_{QT}^2 \times 1/\hat{\beta}_Q = \mathrm{Cov}(Q,\bar{R})/\mathrm{Var}(Q) = \hat{\lambda}$.

Likewise, it can be seen that using separate estimates of the true intake variance σ_T^2 and of the correlation ρ_{QT}^2 for power calculations (based on equation 4) is equivalent to using an estimate of the variance of the predicted intake distribution, $\mathrm{Var}(E[T|Q]) = (\hat{\rho}_{QT}\hat{\sigma}_T)^2$, which can also be obtained directly by the same calibration approach. Thus, when relative risks are expressed for scaled, absolute intake differences, there seems to be no advantage in obtaining separate estimates of β_Q and ρ_{QT} to adjust for bias, nor is there any advantage in separately estimating ρ_{QT} and σ_T^2 for power calculations.

An example

Suppose we wish to investigate whether fat intake as a percent of total energy is associated with breast cancer risk, to confirm observations made in international correlation studies. For this purpose, we plan a prospective cohort study, to be nested within a population-based breast cancer screening programme. Each year, about 12,000 women aged between 50 and 65 visit the screening centre. Every four years, the women are invited for another visit. Thus, within four years, we could recruit a maximum of about 48,000 women in the cohort study. Suppose also that, in this hypothetical population, the expected cumulative incidence of breast cancer is about 200 cases per 10,000 women, after 10 years of follow-up.

For the measurement of dietary intake levels, a self-administered food frequency questionnaire is adopted, which had been previously validated for another on diabetes and cardiovascular disease. Since it is not clear how well the questionnaire will perform in the breast cancer screening population, and to evaluate whether enough variation in fat intake can be measured for the cohort study to reach the required statistical power, it is decided to conduct a one-year pilot study with about 1500 women. Additional objectives of the pilot study are to develop the logistics of the cohort study, and to determine whether the study would not seriously interfere with the normal running of the screening programme. From each study subject, a single reference measurement of dietary intake is obtained using a 24-hour recall interview. Using these pilot study data, the bias factor λ is estimated to be equal to 0.38. The estimated standard deviation of the predicted fat intake distribution is estimated to be 4.2 percent of energy intake (see Table 1).

International correlation studies have reported that, in countries where the diet contains 44 percent calories from fat, breast cancer rates are about 1.5 times higher than in countries where the average fat intake equals 32 percent of energy (20). Assuming that this reflects the true increase in disease risk per percent increase in energy from fat and that, at the level of individuals, the true dose-response relation is exponential, an *a priori* estimate of the log relative risk can be obtained as $\theta_R = \log(1.5)/(44-32) = 0.034$. Thus, using equation 5, the number of cases required for a full cohort analysis (with many more controls than cases) with 0.90 power and 0.05 significance level will be equal to $(1.960+1.282)^2/(0.034\times4.2)^2 \simeq 515$. This number of cases can be reached after following up a cohort of about

Table 1. Measured and predicted distributions of fat intake as a percent of total energy, as estimated in a (pilot) calibration study.

A. Questionnaire measurements:
mean = μ_Q = 37.2
standard deviation = σ_Q = 11.0
B. Reference measurements:
mean = μ_R = 38.5
standard deviation = σ_R = 19.0
Correlation between questionnaire, and reference measurements:
ρ_{QR} = 0.22
Slope of reference measurements regressed on questionnaire measurements
λ = 0.38
C. Predicted intake distribution:
estimated mean = μ_R = 38.5
estimated standard deviation = $\lambda\sigma_Q$ = 4.2

26.000 women for an average of 10 years. Suppose that during the pilot phase about 80 percent of participants in the screening programme agreed to take part in the cohort study. Then, assuming a similar participation rate during the remaining recruitment phase, a sufficiently large cohort could be formed in less than three years.

Discussion

For the efficient planning of prospective cohort studies where diet is the principle exposure factor of interest, it is essential to estimate which is the *minimum study size needed to reach a sufficient level of statistical* power. The estimated sample size requirements, together with the financial resources available, are the key elements to evaluate which detail of exposure information can be obtained from each study participant (e.g.,

including the collection of biologic specimens, or not). Given the true magnitude of increase in disease risk for a standard unit difference in intake level, the power of a cohort study depends on the amount of variation in true intake level that is predicted by the dietary questionnaire measurements collected at baseline. In this paper we have discussed how this predicted variation in intake level can be estimated by a simple calibration approach, using additional reference measurements collected in a sub-sample of cohort participants. The same approach can be used to correct for biases in relative risk estimates. The bias factor (λ) is then equal to the ratio of the predicted variance of the true intake level, divided by the variance of the baseline questionnaire measurements.

As compared to the traditional design of dietary validity studies, a major advantage of the calibration approach is that it requires only a single day's intake record per individual as a reference measurement. Calibration studies can therefore be conducted more easily on a truly representative sub-sample of the study population. Preliminary experience in the European Prospective Investigation into Cancer and Nutrition (EPIC) (17) indicates that almost all study participants will cooperate in 24-hour recall, if this is taken immediately when they present themselves for recruitment in the cohort study.

The proposed calibration approach requires several assumptions:

1) a well defined form of dose-response relation between intake level and incidence rates of disease,
2) normality of the true, as well as of the measured intake distributions, and
3) absence of scaling bias in the reference measurements.

In this paper, we have assumed an exponential type of dose-response. This is a standard model for the analysis of case-control or cohort studies (12,13), which justifies its a priori assumption for statistical power calculations (15,21). Nevertheless, some investigators have based their power calculations on a somewhat different type of risk model. For example, Walker and Blettner (3) and Freudenheim et al. (23) assumed a linear increase in the relative risk (not of its logarithm !) over a series of ordered intake categories, defined by quantiles of the measured intake distribution. Freedman et al. (8) used a similar model, but defined intake categories by quantitative cut-points, on a known (reference) scale, rather

than quantiles. The assumption of a linear trend in relative risk is incompatible with the combination of a normal exposure distribution and a linear logistic model, but differences between these models may be small as long as relative risks are low (as is the case in most studies on diet).

When the cohort study has reached the analysis stage, it can of course be investigated whether a form of dose response provides a better fit of the statistical model to the data observed. Alternatively, a less parametric approach, which is often followed in nutritional epidemiology, is to estimate relative risks for discrete, ordered intake categories, without a priori specification of the form of dose-response. Data from a calibration study can then still be used to estimate the corresponding true mean intake levels within each category. The latter approach is well illustrated by a study on blood pressure in relation to stroke and coronary heart disease (24).

The second assumption implies that not only the individuals' true habitual intake values but also random measurement errors are normally distributed, and that the variance of errors is independent of the level of intake measurements. In practice, this requirement may not always be met. Dietary intake measurements often follow an approximately log-normal distribution, because of larger error variance at higher intake levels. Mathematical transformations can be used to normalize the distribution, and to obtain more constant error variances (25,26). Such transformations, however, also modify the form of dose-response between intake measurements and disease risk, and this should be accounted for in power calculations, or in the estimation of biases in relative risk estimates. This issue requires further investigation. In the meanwhile, it should be realized that normality of the exposure measurements is not only a necessary assumption for the proposed calibration approach, but also for previous methods where the probabilities of misclassification between intake categories are computed from the estimated correlation between measured and true intake values (3,8,23).

The third assumption – absence of scaling biases in the reference measurements – is more specifically required for a valid application of the calibration approach, and puts heavier constraints on the choice of a reference method. For most nutrients, and for food groups, weighed food records or 24-hour recalls (if well conducted) are generally taken to be the optimal methods for measuring mean intake levels of a study population, or

sub-groups (27,28), although the validity of these measurements may be difficult to demonstrate in the absence of objective measures of true intake suitable for use among free-living individuals. Only for a few nutrients, such as protein, is it possible to use a biochemical marker with a well known quantitative relation to absolute daily intake levels (29,30). An apparent advantage of relative risks estimated for quantile categories of intake distribution is that such estimates are independent of the scale on which intake levels are measured. Depending on the variation in intake level within the population, however, a given true increase in disease risk for a standard unit difference in intake level may correspond to different levels of diet-disease association as expressed in relative risks between quantiles. As indicated also by Freedman et al. (8), methods for power calculations based on the a priori assumption of a given relative risk between quantiles (3,23) ignore the amount of variation in intake level present in the study population, and therefore whether the starting assumptions themselves are reasonable. This, in our view, is a greater potential disadvantage than possible under- or overestimation of the predicted intake variance (and thus of statistical power and sample size requirements) which may occur if there is some scaling bias in the reference measurements.

References

1. Willett WC, Sampson L, Stampfer MJ, Rosner B, Bain C, Witschi J, Hennekens CH, Speizer FE. Reproducibility and validity of a semi-quantitative food frequency questionnaire. Am J Epidemiol 1985; 122:51-65.
2. Riboli E, Elmsthål S, Saracci R, Gullberg B, Lindgärde F. The Malmö Food Study: validity of two dietary assessment methods for measuring nutrient intake. Am J Epidemiol (submitted for publication).
3. Walker A, Blettner M. Comparing imperfect measures of exposure. Am J Epidemiol 1985; 121: 783-90.
4. de Klerk NH, English DR, Armstrong BK. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. Int J Epidemiol 1989; 18: 705-12.
5. Willett W. Nutritional epidemiology. New York: Oxford University Press, 1990.
6. Rosner B, Willett B. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. Am J Epidemiol 1988; 127: 377-86.
7. Kaaks R, Riboli E, Estéve J, van Kappel AL, van Staveren WA. Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. Stat Med 1994; 13:127-42.

8. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. Am J Epidemiol 1990; 132:1185-95.
9. Toniolo PG, Pasternack BS, Shore RE, Sonnenschein E, Koenig KL, Rosenberg C, Strax P, Strax S. Endogenous hormones and breast cancer: A prospective cohort study. Breast Cancer Res Treatm 1992; 18:S23-S26.
10. Riboli E, Toniolo P, Shore RE, Kaaks R, Casagrande C, Pasternack BS. Reproducibility of a food frequency questionnaire: effect of self-selection by study subjects. Submitted
11. Rosner B, Willett WC, Spiegelman D. Correction of logistic relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989; 8:1051-69.
12. Breslow NE, Day NE. Statistical methods in cancer research. Volume I. The analysis of case-control studies. IARC Scientific Publications No 32. Lyon: International Agency for Research on Cancer, 1980.
13. Breslow NE, Day NE. Statistical methods in cancer research. Volume II. The design and analysis of cohort studies. IARC Scientific Publications No 82. Lyon: International Agency for Research on Cancer, 1987.
14. Kelsey JL, Thompson WS, Evans AS. Methods in observational epidemiology. New York: Oxford University Press. 1986:254-308.
15. McKeown-Eyssen GE, Thomas DC. Sample size determination in case-control studies: the influence of the distribution of exposure. J Chron Dis 1985; 38: 559-68.
16. Armstrong B, White E, Saracci R. Principles of exposure measurement in epidemiology. Oxford: Oxford Medical Publications. Oxford, 1992:63-4.
17. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer an Nutrition (EPIC). Ann Oncol 1992; 3:783-91.
18. Kaaks R, van der Tweel I, van Noord P, Riboli E. Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential t-test. Epidemiology 1994;5:429-38.
19. Prentice RL, Pepe M, Self SG. Dietary fat and breast cancer: A quantitative assessment of the epidemiological literature and a discussion of methodological issues. Cancer Res 1989;49:3147-56.
20. Schatzkin A, Greenwald P, Byar D, Clifford C. The dietary fat-breast cancer hypothesis is alive. JAMA 1989; 261:3284-7.
21. Wu ML, Whittemore AS, Jung DL. Errors in reported dietary intakes. I. Short-term recall. Am J Epidemiol 1986; 124:826-35.
22. McKeown Eyssen GE, Tibshirani R. Implications of measurement error in exposure for the samples sizes of case-control studies. Am J Epidemiol 1994;139:415-21.
23. Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiological studies of diet and cancer. Nutr Cancer 1988;11:243-50.
24. MacMahon A, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J. Blood pressure, stroke, and coronary disease. Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 1990; 335: 765-74.
25. Box GEP, Cox DR. An analysis of transformations. J Royal Stat Soc, Series B 1964;26:211-52.

26. Borrelli R, Cole TJ, Di Biase G, Contaldo F. Some statistical considerations on dietary assessment methods. Eur J Clin Nutr 1989; 43: 453-63.
27. Cameron ME, van Staveren WA (eds). Manual on methodology for food consumption studies. Oxford: Oxford University Press, 1988.
28. Witschi JC. Short-term dietary recall and recording methods. Guest chapter in: Willett W. Nutritional Epidemiology. New York: Oxford University Press, 1990. pp. 52-68
29. Riboli E, Rönnholm H, Saracci R. Biological markers of diet. Cancer Surv 1987;6:685-718.
30. Hunter D. Biochemical indicators of dietary intake. Guest chapter in: Willett W. Nutritional Epidemiology. New York: Oxford University Press, 1990. pp. 143-216.

Appendix A: Variance of the predicted intake distribution.

If λ is the slope of the linear regression of reference measurements on questionnaire measurements, with an expected value of $\lambda=\rho_{QT}^2/\beta_Q$, and var(Q) is the variance of baseline questionnaire measurements, then the variance of the predicted intake distribution will have an expected value of

$$\mathrm{Var}(E[T|Q]) = \lambda^2\,\mathrm{Var}(Q)$$

$$= [\rho_{QT}^2/\beta_Q]^2\,\mathrm{Var}(Q)$$

$$= 1/\beta_Q^2 \;\frac{(\beta_Q\sigma_T)^4}{(\beta_Q\sigma_T)^2+\sigma_{\varepsilon Q}^2}\;[(\beta_Q\sigma_T)^2+\sigma_{\varepsilon Q}^2]$$

$$= \frac{\beta_Q^2\,\sigma_T^4}{(\beta_Q\sigma_T)^2+\sigma_{\varepsilon Q}^2} = \rho_{QT}^2\,\sigma_T^2$$

Appendix B: The relative precision of λ-estimates.

The variance of the calibration factor estimated in a calibration study of N×k participants with a single reference measurement each, is given by

$$\mathrm{Var}(\hat{\lambda})_1 = [(1-\rho_{QT}^2)\sigma_T^2 + \sigma_{\varepsilon R}^2]/NkVar(Q)$$

$$= [(1-\rho_{QT}^2) + \omega]/Nk \qquad \text{, where } \omega = \sigma_{\varepsilon R}^2/\sigma_T^2.$$

Likewise, the variance of the calibration factor estimated in a calibration study of N individuals, with a k reference measurements each, is given by

$$\mathrm{Var}(\hat{\lambda})_k = [(1-\rho_{QT}^2) + \omega/k]/NVar(Q)$$

The ratio of the two variances equals:

$$\frac{\mathrm{Var}(\hat{\lambda})_k}{\mathrm{Var}(\hat{\lambda})_1} = \frac{k\ [\ (1-\rho_{QT}^2) + \omega/k\ ]}{(1-\rho_{QT}^2) + \omega}\ .$$

It can be seen from this equation that, for all $(1-\rho_{QT}^2) \geq 0$, the variance ratio $\mathrm{Var}(\hat{\lambda})_k/\mathrm{Var}(\hat{\lambda})_1$ will be greater than or equal to 1.0. Thus, for a given total number of reference measurements N, the optimal design of a calibration study nested within a cohort is that with a maximum number of N participants, with only one reference measurement each.

Chapter 4

Adjustments for bias due to errors in exposure assessments in multi-centre cohort studies on diet and cancer: a calibration approach.

This chapter has been published by the authors R. Kaaks, M. Plummer, E. Riboli, J. Estève, and W.A. van Staveren, in the American Journal of Clinical Nutrition (special supplement) 1994;59:245S-50S. The text has been slightly adapted, however, for a more uniform notation of statistical eqations across the chapters of this thesis.

Abstract

An advantage of multi-centre cohort studies on diet and cancer is that these may include populations over a wide range of dietary exposure. With some simplifying assumptions, the information from such multi-centre studies may be divided into:

1. Estimated relationships within each of the separate cohorts, between individual-level measurements of dietary exposure and disease outcome, and
2. An estimated between-cohort relationship, between the mean intake measurements and mean incidence rates.

Errors in the dietary exposure measurements may lead to different amounts of bias in each of these estimated relationships, in particular when dietary questionnaire methods cannot easily be standardized. A calibration approach can be used to adjust for such differences in bias. If sufficiently precise, such calibration will improve the relative weighting of within- , as well as between-cohort components of evidence for a diet-disease association.

Introduction

International correlation studies have shown strong associations between cancer incidence and the per capita intake of specific dietary factors (1). A well known example is the correlation between fat intake and breast cancer incidence. However, because of the possibility of serious uncontrolled confounding, international correlation studies are usually regarded as providing only limited evidence for a causal exposure-disease relationship (2-3). Efforts and resources were therefore focused on studies in which individuals are the units of observation, in the attempt to investigate specific diet-disease relations within populations that are more homogeneous with respect to potential confounding factors (but also with respect to the dietary exposure of interest). An additional methodological advantage of such individual-based studies is that confounding could be adjusted for at the subject level. So far, however, the overall evidence derived from case-control and cohort studies remains inconclusive for many dietary intake factors, some studies showing a significant association while others do not (4) (e.g., total fat intake and breast cancer risk). Inconsistencies between study results could of course be explained by a real absence of relationship between, the dietary factor and disease risk, within some of the populations studied. An alternative interpretation would be that, on average, studies at the subject level lack the statistical power to detect specific diet-disease associations consistently.

A lack of power in case-control or cohort studies on diet and cancer, mostly conducted within limited geographical areas, might be explained to a large extent by the relatively small variations in dietary exposure levels (5). It is reasonable to suppose that true associations between dietary intake factors and cancer risk (if any) are usually relatively weak, with relative risks probably not higher than 4 or 5 for the highest versus the lowest quintile level of intake. Since estimated relative risks are mostly attenuated due to random errors in measurements of exposure, observed diet-disease associations will be even weaker (6,7).

The increase in statistical power that may be obtained within a single study, either by increasing the total sample size, or by improving the precision of the dietary exposure assessments, is often limited for practical or logistical reasons. An additional possibility for improving the statistical power, however, is to increase the overall heterogeneity of the dietary exposure studied, combining the data of multiple studies conducted

in populations with different dietary habits. This was a major underlying rationale for the planning of the European Prospective Investigation into Cancer and Nutrition (EPIC) (8), a collaborative project of multiple cohort studies coordinated by the International Agency for Research on Cancer. In this project, data will be collected on diet and potential confounding or interacting factors such as smoking, physical activity, reproductive history, or drug use. In addition, blood samples will be collected and stored in a biological bank. The study will include about 400,000 middle aged men and women in seven European countries (United Kingdom, Netherlands, Germany, France, Italy, Spain and Greece).

A complicating factor in multi-centre studies such as the EPIC project, is that it may be impossible to use identical dietary assessment methods in each centre. Between countries, food consumption patterns may be as different as the various languages spoken. Since dietary questionnaires should always be adapted to local food habits, they may differ in the number and detail of questions concerning specific foods consumed. Also logistic reasons may preclude the use of identical assessment methodologies. In the EPIC project for instance, it was decided for reasons of compliance to use an interview-administered dietary history questionnaire in southern Italy and Spain. In France, however, most contacts with the subjects will be by mail, necessitating the use of a self-administered (and self-explanatory) food frequency questionnaire.

In this paper we shall describe how, with some simplifying assumptions, the information from multi-centre cohort studies may be considered in two parts:

1. A within-cohort relationship, between intake measurements and disease outcome at the level of individuals, and
2. A between-cohort relationship, between the mean intake measurements and overall incidence rates in the various cohorts.

The estimations of both types of relationship, within and between cohorts, may be biased to different extents as a result of errors in the dietary exposure measurement. This paper discusses the possibility of adjusting for such biases, following a calibration approach. The calibration will improve the relative weighting of the various components of evidence for a diet-disease association, coming from between-subject comparisons within different cohorts, as well as from between-cohort comparisons.

I. Combining cohort-specific relative risk estimates

A model

Consider a collaborative multi-centre investigation of j different cohorts, studying the relation between a given dietary intake factor (e.g., daily fat intake) and the subsequent probability of developing a given disease. Baseline dietary exposure measurements Q are obtained by means of a structured questionnaire. After follow up, in each cohort i (i=1,.....,j) a number of disease cases will be detected, whose dietary intake assessments can then be compared with those of a subset of disease-free individuals. It is then usually assumed that within each cohort there is a linear relation between the logarithm of the disease incidence rate, ψ, and a true underlying exposure T (which may for instance be defined as the average daily fat intake during a given period):

$$\psi_i = \log(\text{rate}) \text{ in cohort } i = \bar{\psi}_i + \theta_i (T - \bar{T}_i)$$

(ignoring, for the sake of simplicity, the effects of confounding or interaction factors). The slopes of these log-linear relations, θ_i, are equal to the logarithm of the relative risk of disease for one unit difference in intake, within a given cohort. The values of θ_i can be estimated for instance by a Poisson type of regression. Alternatively, if the disease incidences are low, but reasonably stable during the period of follow-up, a nested case-control design may be used, and logistic regression may be used to estimate each of these cohort-specific slope parameters (9). Whatever statistical method is used, a crude estimate of log-relative risk, $\hat{\theta}^*$, will be obtained, where the asterisk indicates that the estimate may be biased because it is based on a comparison of questionnaire assessments instead of the true exposure values. Under the mild condition that the exposure distribution is close to normal, and especially under circumstances that the cases are compared to a much larger number of controls, the variance of the θ_i^*-estimate will be approximately equal to (10):

$$\text{Var}(\hat{\theta}_i^*) = \frac{1}{d_i \, \hat{\text{Var}}_i(Q)} \qquad [1]$$

where d_i is the number of cases, and $\text{Var}_i(Q)$ is the variance of the questionnaire measurements of exposure.

Increasing the sample size by combining relative risk estimates across cohorts

Following a meta-analytic approach, the $\hat{\theta}_i^*$-estimates from different cohorts can be combined into an average value $\hat{\theta}_W^*$, summarizing the relations observed within cohorts between exposure measurements and the log incidence rate. Weighting each estimate by the inverse of its variance, this pooled estimate is computed as:

$$\hat{\theta}_W^* = \frac{\sum_{i=1}^{j} d_i \hat{Var}_i(Q) \hat{\theta}_i^*}{D \hat{Var}_W(Q)} \qquad [2],$$

where D is the total number of cases of all cohorts combined, and

$$\hat{Var}_W(Q) = \frac{1}{D} \sum_{i=1}^{j} d_i \hat{Var}_i(Q),$$

is the average within-cohort variance of the questionnaire measurements of exposure (weighted by the numbers of cases in each cohort). The pooled estimate $\hat{\theta}_W^*$ is equivalent to that obtained by, for instance, logistic regression on the data of all cohorts combined, with stratification by cohort. Its variance equals:

$$Var(\hat{\theta}_W^*) = \frac{1}{D \hat{Var}_W(Q)} \qquad [3].$$

Clearly, due to an increase in the number of cases ($D > d_i$), this variance will be smaller than that for any of the cohort-specific $\hat{\theta}^*$-estimates. Thus, if the $\hat{\theta}^*_i$-estimates are approximately equal, indicating a similar trend of dose-response in all cohorts, the efficiency of testing for the presence or absence of a diet-disease association will be improved. This is an obvious advantage of multi-centre studies, whether of a case-control or of a cohort design.

Comparability of cohort-specific relative risk estimates

When combining relative risk estimates into an average summary value, the underlying assumption is that each study provides an estimate of a unique underlying dose-response relationship, which is similar in each of the study populations. Between cohorts, results will then corroborate one another. In practice, however, the study populations of the various cohorts may be quite heterogeneous with regard to the prevalence of additional risk factors. This may cause some between-cohort variation in relative risk estimates, because:

1. to varying degrees, relative risks may be biased due to confounding, and
2. to a different extent, interacting factors may modify the susceptibility to the exposure factor studied (effect modification).

As far as possible, adjustment should be made for confounding by measuring confounders on all individuals in the study, and stratifying the analysis. When in all cohorts individuals are classified into similar strata of age, sex or other potential confounding factors, populations are within such strata more likely to be homogeneous across cohorts. This decreases the likelihood that within similar strata of confounding factors relative risk estimates are differently biased. Classification of individuals into similar strata of confounding factors obviously requires that in each study identical information should be available, not only for the exposure, but also for potential confounding factors. This is a major argument in favour of carefully designed, collaborative multi-centre studies.

The presence of effect modification merely reflects the fact that the relative risk associated with a given type of exposure depends on additional individual characteristics which are unevenly distributed over the various study populations. The assumption that each study provides an estimate of a unique underlying dose response relationship is then violated to at least some degree, and it may not be possible to compute a meaningful summary estimate of relative risk (11); that is, between cohorts study results may not be "combinable" (12). This problem of "combinability" may at least partly be solved if biologically plausible effect modifiers (e.g. smoking, menopausal status, other dietary factors) can be identified at the subject level. If, for example, smoking appears to be a strong effect modifier, between-cohort heterogeneity of results may be explained by differences in smoking prevalence. Results can then be combined across cohorts, but within categories of tobacco consumption. This would require that also information

about potential modifiers be collected in a standard manner in all collaborating centres.

Biases due to errors in the exposure measurements

Biases due to errors in the dietary exposure measurements may form an additional source of between-cohort variation in relative risk estimates. This may be particularly true if it is impossible to use identical dietary assessment methods.

We shall assume that for each cohort the relation between measured and true intakes can be approximately described by a linear measurement error model, which is discussed in more detail elsewhere (13):

$$Q - \bar{Q}_i = \beta_{Qi} (T - \bar{T}_i + \varepsilon_Q) .$$

According to this model, bias in the mean intake measurement, at the group level, is given by the difference $\bar{Q}_i - \bar{T}_i$. The β-coefficient expresses a proportional scaling bias, which occurs if measurement errors are correlated with the true intake values being measured (i.e., when the tendency to over- or under-estimate is different for subjects with a high intake than for subjects with a low intake). The term ε_Q represents an independent, standardized random error, with mean zero and variance $\sigma^2_{\varepsilon Q}$. The ratio of the error variance to the variance of the true exposure (σ^2_T) determines the correlation between measured and true intakes within a given cohort:

$$\rho_{QT,i} = \frac{1}{\sqrt{1 + \sigma^2_{\varepsilon Q}/\sigma^2_T}} \qquad [4].$$

If the measurement error model correctly describes the relation between measured and true intake values, crude estimates of log relative risk ($\hat{\theta}^*$) will be biased by a factor λ which, within a given cohort i, will have an expected value of:

$$\lambda_i = \frac{1}{\beta_{Qi}} \rho^2_{QT,i} \qquad [5].$$

Thus, the bias in the log relative risk estimate will be equal to the inverse of the scaling factor β_Q, multiplied by an attenuation factor which is equal to the square of the correlation between measured and true intakes, ρ^2_{QT}. It can easily be shown that λ is equal also to the slope of a linear regression of true on measured intake values.

Comparisons between dietary validity studies suggest that the correlation between measured and true intake values can for the same nutrient be as different as 0.4 or 0.7, depending on the questionnaire method used and on the study population (14,15). Therefore, λ-values may vary and, even if true relative risks are approximately equal in all cohorts, their estimates may appear to be different. Between-cohort differences in the scaling factor β_Q may add further variability to the estimated relative risk values.

Adjustment for differences in bias due to errors in the exposure measurements: a calibration approach

In order to improve their between-cohort comparability, crude relative risks should be corrected for biases due to measurement error. A convenient approach for making such corrections has been described by Rosner et al, and is referred to as the "linear approximation method" (16). This approach requires that for at least a representative subsample of individuals, there should be an unbiased reference measurement, $R = T + \varepsilon_R$, in addition to the baseline questionnaire assessments (Q). The random errors of the R- and Q-measurements (ε_R and ε_Q, respectively) should be independent. In practice, the reference measurements might for instance be obtained by means of a weighed food record, or using a quantitative biomarker of nutrient intake (if the marker can be reliably translated into an absolute, daily intake value). The bias factor λ can then be estimated in the subsample by normal least squares regression of R- on Q-measurements. Corrected estimates of (log) relative risk are then computed as $\hat{\theta} = \hat{\theta}^*/\hat{\lambda}$. The correction method can also be seen as a regression of disease outcome on "calibrated" dietary questionnaire measurements, which have first been rescaled so that the resulting θ-estimate will be unbiased. We shall therefore also refer to this type of correction as a "calibration".

The variance of the corrected estimates is given as (16):

$$\mathrm{Var}(\hat{\theta}_i) = \frac{1}{\hat{\lambda}_i^2}\,\mathrm{Var}(\hat{\theta}_i^*) + \frac{\hat{\theta}_i^2}{\hat{\lambda}_i^4}\,\mathrm{Var}(\hat{\lambda}_i) \qquad [6].$$

This equation shows that the variance of each corrected θ-estimate will also depend on the precision with which the bias factor λ is estimated. Here, we shall for simplicity assume that the calibration subsample is large enough to obtain λestimates with negligible imprecision (i.e., $\mathrm{Var}(\lambda_i) \approx 0$). The variance of the corrected θestimate then approximately reduces to:

$$\hat{Var}(\hat{\theta}_i) = \frac{1}{\hat{\lambda}_i^2} \hat{Var}(\hat{\theta}_i^*) = \frac{1}{d_i \hat{Var}_i(E[T|Q])} \quad [7],$$

, where

$$Var(E[T|Q]) = \rho_{QT}^2 \sigma_T^2 \quad [8]$$

is equal to the variance of intake predicted by the Q-measurements (i.e., the part of the variance of true intake values T, which is explained in a regression of true intake values on questionaire measurements Q; see Chapter 3, Appendix A).

In this ideal case of perfect calibration, relative risk estimates in each cohort can be accurately adjusted for biases due to errors in exposure measurements. It will thus be possible to to evaluate more accurately whether between-cohort relative risk estimates are in agreement (and "combinable"). Accurate calibration will also result in a more efficient weighting of cohort-specific evidence, when computing an average summary estimate of relative risk. This can be easily seen, since substitution of $Var_i(E[T|Q])$ for $Var_i(Q)$ in equation [2] yields:

$$\hat{\theta}_W = \frac{\sum_{i=1}^{j} d_i \hat{Var}_i(E[T|Q]) \hat{\theta}_i}{D \hat{Var}_W(E[T|Q])} \quad [9].$$

Thus, lower relative weights will be given to cohorts where there is relatively little between-subject variation in true intake values (i.e., when the true intake variance σ_T^2 is small), or where the dietary exposure is relatively poorly measured (i.e., when ρ_{QT} is low). The improved relative weighting of evidence will help optimize the efficiency of a statistical test on the presence of a diet-disease association.

II. Estimates of relative risk based on between-cohort variation of exposure and disease incidence

The corrected summary estimate $\hat{\theta}_W$ is entirely based on comparisons between individuals who belong to the same cohort. An additional estimate of (log) relative risk, $\hat{\theta}_B$, can be obtained on the basis of between-cohort comparisons of disease incidence and exposure levels. Since for both the exposure and incidence the within- and between-cohort components of variance are independent, the estimate of $\hat{\theta}_B$ will be complementary to that of $\hat{\theta}_W$.

As for the within-cohort analysis it was assumed that at the subject level there was a log-linear dose-response between exposure and disease rate, the between-cohort estimate of relative risk should theoretically be obtained by linear regression of estimates of the mean log incidence rate in each cohort , $(\bar{\psi}_i)$, on measurements of the mean exposure levels, $\bar{T}_i$. A complication, however, is that the mean log incidence rates $\bar{\psi}_i$ are not estimable from aggregate level data. (In the literature concerning ecological studies, this has received some attention as a problem of model specification (17,18)). Nevertheless, at least approximate estimates of θ_B can be obtained by linear regression of the logarithm of the average incidence rates, $Y_i=\log(d_i/n_i)$, on measurements of the mean exposures (19,20). This estimate will be most precise if each data point is weighted proportionally to the precision with which it is estimated.

The between-cohort estimate $\hat{\theta}_B$ may be biased if mean exposure measurements $\bar{Q}_i$ are over- or under-estimated to varying degrees. This may be a problem, in particular when the dietary questionnaire methods cannot easily be standardized. One may then choose to obtain an alternative estimate of the mean exposure in only a subsample of each cohort, using an additional dietary assessment method for which standardization poses fewer problems. Again, measurements may for instance be based on a weighed food record or on a biochemical marker. One can thus use the same sub-sample and the same reference measurements (R) as for the within-cohort calibration of relative risk estimates. It should be noted that the reference measurements only need to provide an accurate estimate of the mean exposure at the group level; the measurements may be too imprecise to also provide an accurate classification of individuals by exposure level.

As before, we shall assume that the reference measurements are perfectly standardized, and that calibration sub-samples are sufficiently large to obtain highly accurate estimates of the mean exposure in each cohort. The

precision of each data-point $(\bar{R}_i, Y_i)$ is then mainly determined by the variance of Y_i, which is equal to $1/d_i$. Thus, each point will be optimally weighted by the number of cases, d_i, which leads to the following estimate for θ_B (see also Rothman (21), pp. 336-9) :

$$\hat{\theta}_B = \frac{\Sigma\, d_i\bar{R}_iY_i - \left(\Sigma\, d_i\bar{R}_i\right)\left(\Sigma\, d_iY_i\right) / D}{\Sigma\, d_i\bar{R}_i^2 - \left(\Sigma\, d_i\bar{R}_i\right)^2 / D} = \frac{\hat{Cov}_B(T,Y)}{\hat{Var}_B(T)} \qquad [10].$$

Here, $Cov_B(T,Y) = Cov(E(T|cohort),E(Y|cohort))$ is the estimated between-cohort covariance between the (mean) exposures T_i and estimated average incidence rates Y_i (weighted by the numbers of cases), and $Var_B(T) = Var(E(T|cohort))$ is the between-cohort variance of exposure. The variance of the $\hat{\theta}_B$-estimate is equal to:

$$\hat{Var}(\hat{\theta}_B) = \frac{1}{D\,\hat{Var}_B(T)} \qquad [11]$$

Combining within- and between-cohort estimates of relative risk.

If the within- and between-cohort estimates of relative risk are reasonably similar, both estimates can in principle also be combined, into anoverall summary value $\hat{\theta}_O$. Again, both estimates should be weighted by the inverse of their variances. If both estimates were fully corrected for biases due to error in the exposure measurements (i.e., after perfect calibration), $\hat{\theta}_O$ would be computed as:

$$\hat{\theta}_O = \frac{\hat{Var}_W(E[T|Q])}{\hat{Var}_W(E[T|Q]) + \hat{Var}_B(T)}\,\hat{\theta}_W + \frac{\hat{Var}_B(T)}{\hat{Var}_W(E[T|Q]) + \hat{Var}_B(T)}\,\hat{\theta}_B \qquad [12].$$

Again, the relative weights are proportional to the relative variances, within and between cohorts, of accurately predicted exposure differences. If, on average, exposure differences are poorly distinguished within cohorts (i.e. when the correlation ρ_{QT} is low), the value of $Var_W(E[T|Q])$ decreases, and relatively greater weight will be given to the between-cohort estimate. This illustrates that accurate calibration also improves the weighting of the within-cohort evidence relative to the between-cohort evidence for a

diet-disease association. The variance of the overall $\hat{\theta}_0$-estimate is then equal to :

$$\mathrm{Var}(\hat{\theta}_0) = \frac{1}{D\,(\hat{\mathrm{Var}}_B(T) + \hat{\mathrm{Var}}_W(E[T|Q])} \qquad [13].$$

This formula shows that, compared to the pooled within-cohort estimate $\hat{\theta}_W$, the precision of the relative risk estimate can be further increased by also taking the between-cohort variations in exposure and disease rate into account. Thus, if the estimates $\hat{\theta}_W$ and $\hat{\theta}_B$ are similar enough to be combined, the overall power for testing a single, average dose-response relationship will be improved. This is a major potential advantage specifically for multi-centre cohort studies.

Discussion

A single cohort or case-control study may be unable to detect small increases in relative risk, as it can be the case if the heterogeneity of exposure within a population is small, or when the association between exposure and disease is attenuated by measurement error (or both). This potential limitation of case-control or cohort studies has been used as an argument in favour of ecological studies (19,22-24), since these include a wider range of exposures. However, the disadvantages of ecological studies have also received considerable attention in the epidemiological literature (2,3,18,25). In principle, a collaborative multi-centre cohort study has the ideal design for studying the relationship between dietary intake and disease risk, offering all the potential advantages of studies based on individual subjects (in particular, the possibility to adjust for confounding), while at the same time increasing heterogeneity of the exposure.

With some simplifying assumptions (approximate normality of the exposure distributions, as well as the assumption that the logarithm of the average incidence rate $Y_i=\log(d_i/n_i)$ is a good approximation of the mean log-rate $\bar{\psi}_i$), the information concerning the exposure-disease relationship can be divided into two parts:

1. A within-cohort relationship, between intake measurements and disease outcome at the level of individuals.

2. A between-cohort relationship, between the mean intake measurements and overall incidence rates within each of the various cohorts.

The increased heterogeneity of exposure deriving from a multi-centre design is captured by the data at a cohort level, while within cohorts the evidence is strengthened by an increased number of cases. A similar partitioning of information has been given by Piantadosi et al. (3), as well as by Elliott (26) for the situation where the outcome variable is normally distributed.

The between-cohort part of the study bears a superficial resemblance to an "ecological" study, but is in fact considerably stronger. In true ecological studies, there can be basically two causes of bias (18,25):

a. The population from which exposure data are collected is not representative of the population for which incidence data are available.
b. Adjustment cannot be made for known confounders (and effect modifiers) before aggregation of the data. Attempts to adjust for confounding after aggregation may not be successful.

A multi-cohort study may also suffer from the first of these problems, if there is selection bias in the subsample chosen for standardized exposure measurements. In order to avoid selection bias, one should ensure a very high compliance with the type of dietary method used to obtain the standardized measurements. In the EPIC-project, this was a reason to decide that 24-hour recall interviews will be used as a standard method, rather than weighed food records. The second problem can be overcome in a similar manner as in the analysis of within-cohort relationships, by measuring confounders on all individuals in the study and stratifying the analysis.

Ideally the two forms of evidence - within and between cohorts - should corroborate one another and, following a meta-analytical approach, can ultimately be combined to yield an overall estimate with greater statistical power. The validity of this planned meta-analysis may be increased by careful standardization of measurements of exposure and potential confounding or interacting factors. Correction for biases due to errors in the exposure measurements can be made by a calibration approach, based on the linear approximation method previously described by Rosner et al (16). This will also improve the weighting of component estimates of relative risk (estimated either within cohorts, or from the "ecological", between-cohort relation), the weights being proportional to the variances of the predicted

intake distributions. Likewise, calibration also leads to a reweighting of the overall evidence from within cohorts relative to that from a between-cohort analysis. For the sake of discussion, we made the simplifying assumption that the estimation of calibration parameters (λ) were basically error-free. Sample size requirements for calibration substudies and underlying assumptions regarding the reference measurements will be further discussed elsewhere (20).

Although in principle a collaborative multi-centre cohort study has the ideal design for studying the relationship between dietary intake and disease risk, some caution is necessary. Residual (or unmeasured) confounding may have a different effect on within-cohort relative risk estimates than on estimates based on between-cohort comparisons (25). Another potential problem, which has not been discussed in this paper, is the difficulty of standardizing measurements of the outcome variable. Differences in the completeness of registration of cancer cases, or errors in the evaluation of the time of follow-up may lead to bias in the between-cohort analysis. The interpretation of the study may therefore not be clear whenever there is a strong divergence between the two forms of evidence.

References

1. Armstrong B, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. Int J Cancer 1975; 15: 617-31.
2. Morgenstern H. Uses of ecologic analyses in epidemiologic research. Am J Publ Hlth 1982; 72: 1336-44.
3. Piantadosi S, Byar DP, Green SW. The ecological fallacy. Am J Epidemiol 1988; 127: 893-904
4. Byers T. Diet and cancer. Any progress in the interim? Cancer 1988; 62: 1713-24.
5. Wynder EL, Herbert JR. Homogeneity in nutritional exposure: an impediment in cancer epidemiology. J Natl Cancer Inst 1987; 79: 605-7.
6. Walker AM, Blettner M. Considering imperfect measures of exposure. Am J Epidemiol 1985; 121: 783-90.
7. Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiological studies of diet and cancer. Nutr Cancer 1988; 11: 243-50
8. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992; 3:783-91.
9. Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. J Chron Dis 1983; 36: 715-24.
10. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. J Chron Dis 1967; 20: 511-24.

11. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987; 9: 1-30
12. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987; 316: 450-5.
13. Kaaks R, Riboli E, Estève J, van Kappel AL, van Staveren WA. Estimating the accuracy of dietary questionnaire assessments: Validation in terms of structural equation models. Stat Med 1994; 13:127-42.
14. Pietinen P, Hartman AM, Haapa E, Räsänen l, Haapakoski J, Palmgren J, Albanes D, Virtamo J, Huttunen J. Reproducibility and validity of dietary assessment instruments. II. A qualitative food frequency questionnaire. Am J Epidemiol 1988; 128: 667-76.
15. Tjønneland A, Overvad K, Haraldsdóttir J, Bang S, Ewertz M, Møller-Jensen O. Validation of a semi-quantitative food frequency questionnaire developed in Denmark. Int J Epidemiol 1991; 20: 906-12.
16. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989; 8: 1051-69.
17. Richardson S, Stücker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations. Int J Epidemiol 1987; 16: 111-20.
18. Greenland S. Divergent biases in ecologic and individual-level studies. Stat Med 1992; 11: 1209-23.
19. Prentice RL, Sheppard L. Validity of international, time trend and migrant studies of dietary factors and disease risk. Prev Med 1989; 18: 167-79.
20. Plummer M, Clayton D, Kaaks R. Calibration in multi-centre cohort studies. Int J Epidemiol 1994; in press.
21. Rothman K. Modern epidemiology. Boston, Little, Brown & Co., 1987, pp 336-339.
22. Prentice RL, Sheppard L. Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. Cancer Causes and Control 1990; 1: 81-97.
23. Hiller JE, McMichael AJ. Dietary fat and cancer: a comeback for ecological studies? Cancer Causes and Control 1990; 1: 101-2.
24. Cohen BL. Ecological versus case-control studies for testing a linear-no threshold dose-response relationship. Int J Epidemiol 1990; 19: 680-4.
25. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989; 18: 269-74.
26. Elliott P. Design and analysis of multicentre epidemiological studies: the INTERSALT study. Chapter in: Marmot M, Elliott P. (eds), Coronary heart disease epidemiology, from aetiology to public health. Oxford: Oxford University Press, 1992:168-80.

Chapter 5

Sample size requirements for dietary calibration studies in prospective cohort investigations.

This chapter has been submitted for publication, by the authors R. Kaaks, E. Riboli, and W.A. van Staveren.

Abstract

Advantages of multi-centre cohort studies on diet and cancer is that these allow the cross-validation of relative risk estimates between different study populations. Moreover, more powerful summary estimates of relative risk can be obtained by combining cohort-specific results. A complication, however, is that in different cohorts relative risk estimates may be biased to a different degree as a result of errors in the baseline assessments of habitual dietary intake levels. Such divergent biases can be adjusted for by means of "calibration" studies, using standardized reference measurements obtained in a sub-group of each cohort. These adjustments entail a cost, however, in terms of an increase in the confidence interval the relative risk estimates. In this paper, we evaluate the possible magnitude of such losses in precision, and discuss the approximate sample size requirements of dietary calibration studies for adjustments for bias to have a sufficient level of accuracy.

Introduction

Single case-control or cohort studies may lack the statistical power to detect specific relationships between dietary intake factors and cancer. An advantage of conducting parallel studies in multiple populations is that this allows a cross-evaluation of the consistency of observed diet-disease associations (1). This was one of the reasons for planning the European Prospective Investigation on Cancer and Nutrition (EPIC), a project of multiple cohort studies on diet and cancer in seven European countries (2). An additional advantage of multi-centre cohort studies such as EPIC is that, by combining the data of the different cohorts, the overall study power and precision of relative risk estimates can be improved. A complication, however, is that questionnaire assessments of the individuals' habitual dietary intake levels may not have the same level of accuracy in each of the cohorts. This may be particularly true if it is impossible to use an identical questionnaire in all countries, because of major differences between food consumption patterns or language. Thus, in different cohorts (countries) relative risk estimates may be biased to a varying degree as a result of errors in the dietary exposure assessments, and may not be strictly comparable.

More specifically, suppose that in each study cohort there is a log-linear relation between the incidence rate of disease and the true intake level T of particular nutrient; that is, log(rate) = constant + θT. In this simple model, commonly adopted in the analysis epidemiological data, the parameter θ denotes the logarithm of the relative risk for one unit difference in true intake level. In addition, let us assume that within each cohort the relation between dietary questionnaire assessments and true intake values is well described by a linear measurement error model (3): $Q = \alpha_Q + \beta_Q T + \varepsilon_Q$. Here, the coefficients α_Q and β_Q denote constant and proportional scaling biases, respectively, and the term ε_Q represents a random error with mean zero and variance $\sigma^2_{\varepsilon Q}$. It can then be shown that estimates of the slope parameter θ, obtained by regression of a (binary) disease outcome variable on questionnaire assessments of exposure, will be biased by a factor λ (4,5), with an expected value of

$$\lambda = \frac{1}{\beta_Q} \rho^2_{QT}.$$

Here, ρ_{QT} is the correlation between questionnaire assessments and true intake values. Differences in the proportional scaling bias β_Q, or in the correlation ρ_{QT} between questionnaire assessments and true intake values, will induce different degrees of bias in estimates of the log-relative risk parameter θ, as obtained in each of the various cohorts.

The between-cohort differences in bias can in principle be adjusted for if well standardized reference measurements are available for at least a sub-sample of participants in each cohort. If this adjustment is accurate, it will result in a more optimal weighting of cohort-specific relative risk estimates when these are combined into a summary value, the weighting factors reflecting the level of accuracy of the dietary intake assessments in each cohort (6). The reference measurements may be obtained by a more detailed method, such as daily intake records, using a weighing method or 24-hour diet recalls (7,8). Then, using the method of "linear approximation" described by Rosner et al. (4), the bias factor λ can in each cohort be estimated by normal linear regression of reference measurements on questionnaire assessments, and corrected estimates of the log-relative risk can be obtained as $\hat{\theta}=\hat{\theta}^*/\hat{\lambda}$ (where $\hat{\theta}^*$ is the unadjusted estimate). This method can also be seen as a rescaling of the baseline questionnaire assessments so that, by regression of disease outcome on the rescaled assessments, a consistent estimate of the log-relative risk will be obtained (6). We shall therefore refer to this approach as a "calibration" procedure, and to the factor λ as the "calibration factor".

Adjustments of relative risk estimates by calibration will only be adequate if, in each cohort, the calibration factor λ is estimated with a sufficient level of precision. In this paper, we present some mathematical simulations to illustrate the possible losses in precision which may result from calibration. In the subsequent sections we shall discuss the approximate sample size requirements of dietary calibration studies, in prospective cohort studies on diet and chronic disease risk.

Loss of precision due to calibration

Rosner et al. (4) derived a closed-form expression for the variance for the calibrated estimate of the log-relative risk ($\hat{\theta}=\hat{\theta}^*/\hat{\lambda}$), taking account of the imprecision in the estimation of the calibration factor λ. We shall assume that the exposure distribution is close to normal, with mean μ_T and variance σ^2_T, and that the cases are compared to a much larger number of disease-free

control subjects as in the case of a full cohort analysis. The variance formula by Rosner et al. can then be re-expressed in terms of an "effective" number of cases, $\tilde{D}$, and of the amount of variation in true intake level actually predicted by the questionnaire assessments, $\mathrm{Var}(E[T|Q]) = \rho_{QT}^2 \sigma_T^2$:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{\tilde{D}} \frac{1}{\mathrm{Var}(E[T|Q])} \qquad [1]$$

(Appendix A). Here, the effective number of cases $\tilde{D}$ is given by

$$\frac{1}{\tilde{D}} = \frac{1}{D} + C \qquad [2]$$

where $C = \frac{1}{N} A^2 \frac{1 - \rho_{QR}^2}{\rho_{QR}^2}$, and $A^2 = \theta^2 \mathrm{Var}(E[T|Q])$.

In this equation, N denotes the number of participants in the calibration sub-study; and ρ_{QR} is the correlation between questionnaire and reference measurements. The quantity A indicates the strength of the association between questionnaire measurements and disease risk, and can be directly related to an expected relative risk between quantile levels of the distribution of questionnaire assessments. For instance, the relative risk for the upper versus the lower quintile of the questionnaire assessments approximately equals $e^{2.80A}$ (9), which corresponds to a relative risk of about 1.5 when the association A equals 0.15, or a relative risk of 4.0 when the association equals 0.50.

The quantity C in equation [2] will be called the statistical "cost" of the calibration (not to be confounded with the financial cost of the calibration study). A cost equal to zero corresponds to estimation of the calibration factor λ without error and, therefore, to no increase in the variance of the calibrated estimate of the log-relative risk. In this hypothetical situation, the effective number of cases is equal to the number actually observed in the study (i.e., $\tilde{D}=D$). In practical situations, however, the calibration factor will be estimated with some level of

imprecision. In this case, the ratio $\tilde{D}/D$ expresses the relative efficiency of the calibration study, as compared to the situation where the calibration is perfect. The loss of efficiency can be defined as $D-\tilde{D}$.

To illustrate the potential magnitude of such efficiency losses, Table 1 shows the effective numbers of cases at selected values of the association between questionnaire assessments and disease risk (A), and of the observed number of cases (D). The relative efficiency improves with an increasing number of subjects in the calibration study, or with increasing values for the correlation between questionnaire and reference measurements. The latter two parameters determine the precision with which the calibration factor λ is estimated. On the other hand, the relative efficiency decreases with increasing values of the association A, or of the observed number of cases D, which both determine the precision of the crude log-relative risk estimate before calibration.

An additional parameter of interest (shown between parentheses in Table 1) is the expected value of the calibrated estimate of log-relative risk divided by its standard error – that is, the expected t-value. This expected t-value is determined by the association between questionnaire assessments and disease risk, as well as by the effective number of cases, and can be computed as $E(t)=A\sqrt{\tilde{D}}$ (see Appendix B). If the 95 percent confidence interval of the calibrated log-relative risk estimate is expressed as a fraction f of the estimate itself, it can be shown that $f=1.96/E(t)$ (Appendix B)). Thus, an expected t-value of 3.92 ($\simeq$4.0) corresponds to a confidence interval of $[1\pm0.50]\theta$.

Efficient design of calibration studies: number of repeat reference measurements

In practice, the statistical cost of calibration, C, can be reduced in two ways: increasing the number of subjects in the calibration study, or taking the average of multiple reference measurements per subject (e.g., multiple food records, or 24-hour recalls) to strengthen their correlation (ρ_{QR}) with the questionnaire measurements. The question is which of these two approaches will reduce the cost of calibration most efficiently.

The statistical cost for a calibration study of N=n study participants with reference measurements repeated over k days ($C_{k>1}$), relative to that for a study of N=n×k participants with only a single reference measurement ($C_{k=1}$), can be computed as:

Table 1. Effective numbers of cases, and (between parentheses) expected t-values after calibration.

a) The expected odds ratio for the upper versus the lower quintile of the measured exposure distribution equals 1.50 (i.e., A=0.15).

	Number of participants in calibration study, N					
	1000	2000	3000	4000	5000	∞
Nr. observed cases, D = 100						
ρ_{QR} = 0.1	81.8 (1.36)	90.0 (1.42)	93.1 (1.45)	94.7 (1.46)	95.7 (1.47)	100 (1.50)
ρ_{QR} = 0.2	94.9 (1.46)	97.4 (1.48)	98.2 (1.49)	98.7 (1.49)	98.9 (1.49)	100 (1.50)
ρ_{QR} = 0.3	97.8 (1.48)	98.9 (1.49)	99.2 (1.49)	99.4 (1.50)	99.5 (1.50)	100 (1.50)
ρ_{QR} = 0.4	98.8 (1.49)	99.4 (1.50)	99.6 (1.50)	99.7 (1.50)	99.8 (1.50)	100 (1.50)
ρ_{QR} = 0.5	99.3 (1.49)	99.7 (1.50)	99.8 (1.50)	99.8 (1.50)	99.9 (1.50)	100 (1.50)
Nr. observed cases, D = 500						
ρ_{QR} = 0.1	237 (2.31)	321 (2.69)	365 (2.87)	391 (2.97)	409 (3.03)	500 (3.35)
ρ_{QR} = 0.2	394 (2.98)	441 (3.15)	459 (3.21)	468 (3.24)	474 (3.27)	500 (3.35)
ρ_{QR} = 0.3	449 (3.18)	473 (3.26)	482 (3.29)	486 (3.31)	489 (3.32)	500 (3.35)
ρ_{QR} = 0.4	472 (3.25)	486 (3.31)	490 (3.32)	492 (3.33)	494 (3.33)	500 (3.35)
ρ_{QR} = 0.5	484 (3.30)	492 (3.33)	494 (3.33)	496 (3.34)	497 (3.34)	500 (3.35)

Table 1 (continued)

b) The expected odds ratio for the upper versus the lower quintile of questionnaire measurements equals 4.00 (i.e., A=0.50).

	Number of participants in calibration study, N					
	1000	2000	3000	4000	5000	∞
Nr. observed cases, D = 100						
ρ_{QR} = 0.1	28.8 (2.68)	44.7 (3.34)	54.8 (3.70)	61.8 (3.93)	66.9 (4.09)	100 (5.00)
ρ_{QR} = 0.2	62.5 (3.95)	76.9 (4.38)	83.3 (4.56)	87.0 (4.66)	89.3 (4.72)	100 (5.00)
ρ_{QR} = 0.3	79.8 (4.47)	88.8 (4.71)	92.2 (4.80)	94.06 (4.85)	95.2 (4.88)	100 (5.00)
ρ_{QR} = 0.4	88.4 (4.70)	93.8 (4.84)	95.8 (4.89)	96.8 (4.92)	97.4 (4.93)	100 (5.00)
ρ_{QR} = 0.5	93.0 (4.82)	96.4 (4.91)	97.6 (4.93)	98.2 (4.95)	98.5 (4.96)	100 (5.00)
Nr. observed cases, D = 500						
ρ_{QR} = 0.1	37.4 (3.06)	69.6 (4.17)	97.6 (4.94)	122 (5.52)	144 (6.00)	500 (11.2)
ρ_{QR} = 0.2	125 (5.59)	200 (7.07)	250 (7.91)	286 (8.46)	313 (8.85)	500 (11.2)
ρ_{QR} = 0.3	221 (7.43)	306 (8.75)	352 (9.38)	380 (9.75)	399 (9.99)	500 (11.2)
ρ_{QR} = 0.4	302 (8.69)	376 (9.69)	410 (10.1)	430 (10.4)	442 (10.5)	500 (11.2)
ρ_{QR} = 0.5	364 (9.54)	421 (10.3)	444 (10.5)	457 (10.7)	465 (10.8)	500 (11.2)

$$\text{"Relative Cost"} = \frac{C_{k>1}}{C_{k=1}} = \frac{k\,[(1-\rho_{QT}^2) + (1-\gamma)\omega/k]}{(1-\rho_{QT}^2) + \omega} \qquad [3]$$

where $\omega = [1-\rho_{RT}^2]/\rho_{RT}^2$, and where γ is the correlation between random errors of repeat reference measurements taken on the same individual (Appendix C). It can be immediately seen from this equation that, unless there is a perfect correlation between questionnaire assessments and true intake values (i.e., unless $(1-\rho_{QT}^2)=0$), the "relative cost" of calibration is always greater than 1.0. Thus, for a given total number of reference measurements, the calibration study will be most efficient if it includes a maximum number of subjects with only a single measurement each (i.e, k=1).

Table 2.a shows the relative cost of calibration for different numbers of repeat reference measurements per subject, at selected values of the correlations ρ_{QT} and ρ_{RT}, assuming that errors of repeat reference measurements are uncorrelated. Likewise, table 2.b shows the relative cost when there is a 0.10 correlation between the random errors of repeat reference measurements. From the two tables it can be seen that, with 14 or more repeats (which is a common number in dietary validity studies (10,11)), the relative cost can be higher than 2.0, in particular when random errors of repeat reference measurements are not fully independent. A relative cost of 2.0 means that a similar precision of calibration can be attained with half the total number of reference measurements if, instead of taking replicate measurements, more subjects are included in the study.

Precision of calibration studies: sample size requirements

Given our conclusion that the cost for calibration C is more efficiently reduced by increasing the total sample size N, rather than by taking replicate reference measurements, the main question to be answered is which sample size will be needed for the calibration study to reach a desired level of precision.

Ideally, the calibration study should be as large as to maintain a minimum level of relative efficiency $\tilde{D}/D$ of, say, 80 percent. From equation [2] we can derive,

TABLE 2. "Relative Cost" for calibration with k repeat reference measurements, or only a one reference measurement per participant, assuming a 0.50 correlation between questionnaire assessments and true intake values (i.e., ρ_{QT}=0.50).

a) Zero correlation between errors of replicate reference measurements.

	correlation between reference measurements and true intake level, ρ_{RT}				
	0.10	0.20	0.30	0.40	0.50
Nr of repeat measurements, k					
2	1.01	1.03	1.07	1.13	1.20
7	1.05	1.18	1.41	1.75	2.20
14	1.10	1.39	1.90	2.62	3.60

b) Correlation (γ) between errors of replicate reference measurements equals 0.10.

	correlation between reference measurements and true intake level, ρ_{RT}				
	0.10	0.20	0.30	0.40	0.50
Nr of repeat measurements, k					
2	1.10	1.13	1.16	1.21	1.28
7	1.64	1.76	1.97	2.28	2.68
14	2.39	2.65	3.11	3.76	4.64

$$\frac{D}{\tilde{D}} = 1 + A^2 \frac{1 - \rho_{QR}^2}{\rho_{QR}^2} \frac{D}{N} \qquad [4]$$

Thus, at any given value of the association A and of the correlation ρ_{QR}^2 between questionnaire assessments and reference measurements, there is a positive relation between the relative efficiency $\tilde{D}/D$ and the ratio N/D which is the size of the calibration study relative to the observed number of cases. For example, assume that the correlation between questionnaire and reference measurements equals ρ_{QR}=0.20. (This was the level of correlation between intake assessments obtained by a questionnaire, and reference measurements obtained by a single 24-hour recall, as observed for most nutrients during the pilot phase of the EPIC study (personal communication).) In addition, assume that the relative risk between the extreme quintiles of questionnaire assessments equals 1.50 (i.e., A=0.15). It then follows that the calibration study should be about 2.2 times as large as the number of cases observed.

Unfortunately, using the relative efficiency $\tilde{D}/D$ as the main criterion, sample size requirements for the calibration study may become excessively large when the association between questionnaire assessments and disease risk is comparatively strong. For example, if the relative risk between the extreme quintiles equals 4.0 (i.e., A=0.50), the calibration study should be 24 times larger than the number of cases observed. Thus, if 500 cases with disease are expected within the cohort, the calibration study should include 12,000 participants! One may argue however that, when the unadjusted log-relative risk estimate has very narrow confidence limits, the confidence interval can still have an acceptable width even in situations where the relative efficiency of calibration is less than 80 percent. Thus, when there is a strong association between questionnaire assessments and disease risk, an alternative may be to use the expected t-value for the calibrated log-relative risk estimate as a criterion for sample size calculations .

From equation [2], we can derive an inverse relation between the expected t-value of the calibrated log-relative risk estimate, and the relative efficiency of the calibration study:

$$E(t) = \sqrt{\{ (1 - \tilde{D}/D)\ N\ \rho_{QR}^2/[1-\rho_{QR}^2] \}} \qquad [5].$$

Using this relationship, and making an assumption only about the correlation between questionnaire and reference measurements, a minimum sample size for the calibration study can be computed so that either

A. a minimum level of relative efficiency will be attained (e.g., $\bar{D}/D \geq 0.80$), in which case the expected t-value would be only marginally greater even if the calibration was perfect, or
B. the expected value of the calibrated log-relative risk estimate divided by its standard error (i.e., the expected t-value) reaches a desired minimum value.

For example, assuming that the correlation ρ_{QR} is greater than or equal to 0.20, it follows from equation [5] that the calibration study should include at least 1920 (≈2000) subjects to obtain that either the expected t-value is greater than 4.0, or the relative efficiency $\tilde{D}/D$ is greater than 0.80.

Discussion.

An advantage of multi-centre cohort studies on diet is that these allow the consistency of relative risk estimates across populations to be evaluated. Moreover, by combining relative risk estimates across cohorts, a more powerful summary estimate can be obtained to test for specific diet-disease associations. A complication, however, is that errors in the baseline questionnaire assessments of dietary intake levels may lead to different degrees of bias in each of the cohort-specific estimates of relative risk. Adjustments for such divergent biases can be made by means of dietary calibration studies, using well standardized reference measurements collected within a sub-sample of each cohort. In this paper we have evaluated the potential loss in precision of relative risk estimates, within a single cohort study, as a result of such calibration adjustments.

Within a single study population (i.e., within each cohort separately), a test for association between a given dietary factor and disease risk (i.e., a test for the null hypothesis that $\theta=0$) has been shown to have optimal power if it is based on the unadjusted log-relative risk estimate (12). On the other hand, when multiple relative risk values are combined into a single summary value, calibration can remove error by adjusting for heterogeneity caused by dietary assessment errors. In a multi-centre study, therefore, calibration can in principle also improve the precision of a

combined log-relative risk estimate, provided that this potential gain in precision is larger than the intra-cohort losses of efficiency ($\tilde{D}/D$). We have shown how minimum sample size requirements for calibration studies can be computed so that

1) the relative loss in efficiency will be small; this will be the case mainly if the association between questionnaire assessments and disease risk is relatively weak or if the number of cases is small; or
2) the expected t-value of the calibrated log-relative risk estimate reaches a minimum predefined value; this will occur only if the unadjusted estimate reaches a certain level of precision – that is, when the association between questionnaire assessments and disease risk is relatively strong, or when there is a large number of cases.

It is in particular in the first type of situation, when there are only weak associations between questionnaire assessments and disease risk, that the benefit of reducing error due to unequal biases in relative risk estimates may be relevant, to optimize the statistical power.

Our calculations were based on Rosner et al.'s variance formula for the calibrated log-relative risk estimate (4). Using computer simulations, Rosner et al. have shown that the linear approximation method results in a satisfactory reduction of bias, as long as true relative risks are relatively low (as is the case in most studies on diet and cancer). In addition, it was shown that confidence intervals for the corrected log-relative risk estimate will have a probability of covering the true log-relative risk value that is very close, although not identical, to the nominal (95 percent) level. The variance formula for the calibrated log-relative risk estimate was derived by Rosner et al. on the assumption that the calibration study would be conducted externally from the main epidemiological study population. In prospective cohort studies, however, calibration studies can be conducted on a random sub-sample of study participants. Thus, for a small proportion of the cohort (e.g., in 2000 out of 50,000 participants), reference measurements of dietary exposure will be available, which provide some additional information on the subjects' classification by habitual intake level. Since some cases with disease may arise in this sub-sample, the calibration studies may actually provide some supplementary information about the exposure-disease relationship. By ignoring this supplementary information, efficiency losses and sample size

requirements may have been somewhat overestimated. Nevertheless, this overestimation will be negligible as long as the expected number of cases in the sub-sample is low and, in addition, if the reference measurements (e.g., based on only a single 24-hour recall) have a comparatively low reliability as assessments of the individuals' long-term, habitual intake levels.

Mathematically more exact, likelihood-based approaches for the calculation of the sample size requirements of dietary calibration studies have been developed, which do take account of additional exposure information for cases and controls within the calibration sub-sample (13,14). These approaches allow a simultaneous evaluation of sample size requirements of the main cohort, as well as of the proportion of the cohort that to be allocated to a calibration sub-study. The relative efficiencies of calibration, and expected t-values as presented in this paper are based on simplified formulas. In situations where the size of the main study cohort has already been decided, however, these parameters can provide simple and practical criteria for the estimation of sample size requirements for dietary calibration studies.

References

1. Friedenreich CM, Brant RF, Riboli E. Influence of methodologic factors in a pooled analysis of 13 case-control studies of colorectal cancer and dietary fiber. Epidemiology, 1994; 5:66-79.
2. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992; 3:783-91.
3. Kaaks R, Riboli E, Estève J, van Kappel AL, van Staveren WA. Estimating the accuracy of dietary questionaire assessments: Validation in terms of structural equation models. Stat Med 1994; 13:127-42.
4. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989; 8: 1051-69.
5. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate error: application to diet and colon cancer. Stat Med 1989; 8: 1151-63.
6. Kaaks R, Plummer M, Riboli E, Estève J, van Staveren WA. Adjustment for bias due to errors in exposure assessments in multi-centre cohort studies on diet and cancer: a calibration approach. Am J Clin Nutr, 1994; 59(suppl):245S-50S.
7. Cameron ME, van Staveren WA (eds). Manual on methodology for food consumption studies. Oxford: Oxford University Press, 1988.
8. Witschi JC. Short-term recall and recording methods. Guest Chapter in: Willett WC. Nutritional Epidemiology. New York: Oxford University Press, 1990:53-68.

9. Kaaks R, van der Tweel I, van Noord P, Riboli E. Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential t-test. Epidemiology 1994;5:429-38.
10. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semi-quantitative food frequency questionnaire. Am J Epidemiol 1985; 122: 51-65.
11. Pietinen P, Hartmann AM, Haapa E, Räsänen L, Haapakoski J, Palmgren J, Albanes D, Virtamo J, Huttunen J. Reproducibility and validity of dietary assessment instruments. II. A qualitative food frequency questionnaire. Am J Epidemiol 1988; 128: 667-76.
12. Tosteson TD, Tsiatis AA. The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. Biometrika 1988; 75:507-14.
13. Sullivan Pepe M, Self S, Prentice RL. Further results on covariate measurement errors in cohort studies with time to response data. Stat Med 1989; 8:1167-78.
14. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics 1991; 47: 851-69.

Appendix A

It was shown by Rosner, Willet and Spiegelman (4) that the variance of the calibrated θ-estimate, Var(θ), will be approximately equal to:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{\hat{\lambda}^2}\ \mathrm{Var}(\hat{\theta}^*) + \frac{\hat{\theta}^{*2}}{\hat{\lambda}^4}\ \mathrm{Var}(\hat{\lambda}) \qquad [A.1]$$

Assuming that intake levels are Normally distributed, and that cases are compared to a much larger number of controls (as in a full cohort analysis), the variance of the unadjusted θ-estimate ($\hat{\theta}^*$) is equal to

$$\mathrm{Var}(\hat{\theta}^*) = \frac{1}{D\ \mathrm{Var}(Q)} \qquad [A.2]$$

Furthermore,

$$\mathrm{Var}(\hat{\lambda}) = \frac{\mathrm{Var}(R)\ [1-\rho_{QR}^2]}{N\ \mathrm{Var}(Q)} \qquad [A.3]$$

$$E(\hat{\lambda}) = \frac{1}{\beta_Q}\ \rho_{QT}^2 \qquad [A.4]$$

Substitution of [A.2], [A.3], and [A.4] into equation [A.1], and

reexpressing

$\mathrm{Var}(R) = \sigma_T^2/\rho_{RT}^2$ and $\mathrm{Var}(Q) = \beta_Q^2\,\sigma_T^2/\rho_{QT}^2$ yields:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{\rho_{QT}^2\,\sigma_T^2}\left(\frac{1}{D} + \frac{\theta^2\,\sigma_T^2}{N\,\rho_{RT}^2}\,[1-\rho_{QR}^2]\right)$$

$$= \frac{1}{\rho_{QT}^2\,\sigma_T^2}\left(\frac{1}{D} + \frac{1}{N}\,\theta^2\,\sigma_T^2\,\rho_{QT}^2\,\frac{1-\rho_{QR}^2}{\rho_{QR}^2}\right)$$

$$= \frac{1}{\rho_{QT}^2\,\sigma_T^2}\,\frac{1}{\tilde{D}} = \frac{1}{\mathrm{Var}(E[T|Q])}\,\frac{1}{\tilde{D}} \qquad [A.5]$$

Appendix B

The expected t-statistic for a test whether the calibrated θ-estimate significantly differs from 0 equals:

$$E(t) = \frac{E(\hat{\theta})}{\sqrt{\mathrm{Var}(\hat{\theta})}}$$

$$= \frac{E(\hat{\theta})}{1/\sqrt{(\tilde{D}\ \mathrm{Var}(E[T|Q])}}$$

$$= \theta\,\sqrt{\mathrm{Var}(E[T|Q])}\,\sqrt{\tilde{D}}$$

$$= A\,\sqrt{\tilde{D}} \qquad [B.1].$$

Suppose the expected 95% confidence interval for the calibrated θ-estimate equals $(1\pm f)\theta$. Then, $f\theta = 1.96\ \sqrt{\mathrm{Var}(\theta)}$, and thus $f = 1.96\ \sqrt{\mathrm{Var}(\theta)}/\theta = 1.96/E(t)$, where 1.96 is the 0.95 point of the standard Normal distribution.

Appendix C.

The relative cost R.C., for a calibration study of N=n subjects each with k repeat reference measurements, relative to that for a study of N=n×k subjects with only a single reference measurement each, equals

$$R.C. = \frac{1/n \;\; A^2 \;\; [1-\rho^2_{QR,k}]/\rho^2_{QR,k}}{1/(nk) \;\; A^2 \;\; [1-\rho^2_{QR}]/\rho^2_{QR}} =$$

$$= \frac{k \; [\; 1/\rho^2_{RT,k} \; - \; \rho^2_{QT} \;]}{1/\rho^2_{RT} \; - \; \rho^2_{QT}} \qquad [C.1].$$

Here, ρ_{QR} is the correlation between the questionnaire, and (single) reference measurements, which can be rewritten as

$$\rho_{QR} = \rho_{QT}\rho_{RT} = \rho_{QT} \; 1/\sqrt{\{1+\omega\}} \qquad [C.2].$$

where ω is the variance ratio $\omega=\sigma^2_{\varepsilon R}/\sigma^2_T$.

Likewise, $\rho_{QR,k}$ is the correlation between questionnaire measurements and the average of k repeat reference measurements, and can be rewritten as

$$\rho_{QR,k} = \rho_{QT}\rho_{RT,k} = 1/\{1+\gamma\omega+(1-\gamma)\omega/k\} \qquad [C.3],$$

where γ is the correlation between random errors of replicate reference measurements.

Substitution of [C.2] and [C.3] into [C.1] yields

$$R.C. = \frac{k \; [\; (1-\rho^2_{QT}) \; + \; \gamma\omega \; + \; (1-\gamma)\omega/k \;]}{(1 \; - \; \rho^2_{QT}) \; + \; \omega} \qquad [C.4].$$

Chapter 6

Efficient use of biological banks for biochemical epidemiology: exploratory hypothesis testing by means of a sequential t-test.

This chapter has been published by the authors R. Kaaks, I. van der Tweel, P.A.H. van Noord, and E. Riboli, in Epidemiology 1994;5:429-38.

Abstract

In view of recent advances in molecular and biochemical epidemiology, there is growing interest in the creation of biological banks of blood, urine, tissue or other biological specimens collected from participants in prospective cohort studies. The existence of biological banks may make it possible to study a multitude of etiological hypotheses, by comparing biochemical parameters measured in the biological specimens of subjects who will eventually develop the disease of interest ("cases") and of control subjects, using a nested case-control or a case-cohort design. In practice, however, the amount of biological material available per subject (in particular that of cases) will limit the number of hypotheses that can be tested. The present paper discusses the use of a sequential t-test, which, compared to an analogous fixed-sample procedure, will on average require fewer biological specimens to accept or reject a given study hypothesis. The sequential test may thus facilitate an early decision on whether a new hypothesis is worth further investigation, while avoiding to use too much biological material on testing hypotheses that would eventually prove unfruitful. If the test reveals an exposure difference of interest, the study may be extended so as to allow more accurate estimation of relevant epidemiological effect measures.

Introduction

Following recent developments in "biochemical" and "molecular" epidemiology there is growing interest in the creation of banks of biological samples of material, such as blood or urine specimens, collected from participants in prospective cohort studies (1,2). After detection of a sufficient number of cases of a given disease (during a given follow-up period), parameters measured in their biological specimens can be compared with those of controls in order to study specific etiological hypotheses. Since new laboratory techniques are constantly being developed for the assessment of specific biochemical or molecular parameters, the number of new hypotheses that can be tested is also increasing rapidly. In practice, however, the amount of biological material stored (in particular that of cases) will limit the number of possible studies (3). It would therefore be useful to have a statistical method which, at the expense of as little biological material as possible, allows a distinction between promising hypotheses, which may be worth further investigation, and less promising ones. Such a method may be particularly useful in exploratory investigations, when there is only limited prior evidence to justify a study based on a large number of biological specimens.

Using sequential statistical designs (4,5), it is theoretically possible to terminate an investigation on a specific hypothesis as soon as sufficient evidence has accumulated for it to be accepted or rejected. On average, sequential analysis will arrive at a decision after substantially fewer observations than equally reliable test procedures based on a fixed sample size. The first sequential procedures were developed during the Second World War (6), when Wald described the theoretical basis for a sequential probability ratio test (SPRT), and it almost immediately became an important tool for efficient quality control in wartime factories. Nowadays, sequential methods have also been adopted for use in medical research, in particular for the design and analysis of clinical trials (7,8). So far, however, sequential methods have not been much used in epidemiological studies, outside clinical trials.

The present paper discusses exploratory hypothesis testing by means of a sequential t-test, in cohort-nested case-control studies where the exposure assessment is based on a biochemical marker, obtained by laboratory analysis of stored biological specimens. (To simplify, we shall refer to the biological marker as a measurement of an internal or external "exposure",

although it is clear that markers can also be a measure of individual susceptibility or of intermediate endpoints (9,10)). The application of the sequential t-test will be illustrated using data from a study conducted to examine whether selenium is a potentially protective agent against breast cancer (11).

The sequential t-test

We shall assume that the biomarker measurements, M, can be considered as values drawn from two normal distributions, for cases and for controls, respectively. We also assume that both distributions have an equal variance, σ^2, but that their means may be different; that is:

$M|\text{case} \simeq N(\mu_1,\sigma^2)$, and

$M|\text{control} \simeq N(\mu_0,\sigma^2)$.

The null hypothesis to be tested is that the mean exposures of cases and controls are equal; that is,

$H_0 : \mu_1=\mu_0$, or $\mu_1-\mu_0=0$.

If σ is not known a priori, but must be estimated, the magnitude of the mean difference $\mu_1-\mu_0$ which can be detected with a given power is unknown. However, the null hypothesis can be re-defined in terms of a standardized difference, $\theta=(\mu_1-\mu_0)/\sigma$, between the mean exposures of cases and of controls:

$H_0 : \theta = (\mu_1-\mu_0)/\sigma = 0$.

If the standard deviation σ is high, then for a given number of observations only very large differences will be detectable with sufficient statistical power. Inversely, the power will be higher if σ is small.

A t-test can be used to evaluate the null hypothesis against an alternative. In the case of a well defined biological hypothesis, a one-sided alternative may be reasonable; that is,

$H_1 : \theta \geq \theta_R$.

Here, θ_R is the minimum standardized difference $(\mu_1-\mu_0)/\sigma$ [1] that one would find relevant enough to be detected, with a power of at least $1-\beta$ and a significance level α. A two-sided alternative can be specified as

$H_1 : |\theta| \geq \theta_R$.

Most epidemiologists are familiar with the traditional, fixed sample t-test,

[1] Note: If the exposure is expected to be higher for controls than for cases, the standardized difference can also be defined as $\theta = (\mu_0-\mu_1)/\sigma$.

based on the comparison of the mean exposures of predetermined numbers of cases and controls. The procedure described here, however, uses a sequential sampling of cases and controls within the cohort. This sequential sampling may follow the detection of cases over time. Alternatively, if a large number of cases has already accrued, the sequential sampling can also be performed retrospectively. In the latter situation, the order in which cases are selected does not need to follow the chronological order in which they were detected, but can also be based on a random selection process. For each case selected, a random subset of k controls is drawn from the disease-free subjects in the cohort. If there are many cases, and if the major concern is to limit the additional costs for laboratory analyses, 1:1 matching (k=1) will give optimal statistical power at a given total cost. However, when disease incidence rates are low (e.g., for a given type of cancer), cohort studies must be very large to observe a sufficient number of cases. Additional costs for laboratory assessments – even though considerable – may then still be low in comparison to the initial investments in the study, and priority may be given to the possibility of studying as many hypotheses as possible with the biological material available. In this case, a higher matching ratio will be more efficient (k>1), as this will increase the power of the test keeping the number of specimens from cases constant. A matching ratio greater than 5 will seldom be worthwile, however (12). After every new set of one case plus corresponding controls is sampled, the biochemical measurements are compared for all cases and controls processed up to that point to determine whether there is sufficient evidence to either reject or accept the null hypothesis H_0.

The earliest theory for sequential test procedures (that of the sequential probability ratio test), was initially developed by Wald (6). According to this theory, a sequential test was based on the logarithm of the following likelihood ratio, L_n, which can be computed after every new case-control set is being sampled:

$$L_n = \frac{\text{the probability of observing the case and control measurements if } H_1 \text{ is true (i.e., if } \theta \geq \theta_R)}{\text{the probability of observing the case and control measurements if } H_0 \text{ is true (i.e., if } \theta = 0)}.$$

(where n is the number of case-control sets processed so far). A high value of the logarithm of the likelihood ratio, l_n, indicates that, given the measurements observed, the alternative hypothesis H_1 is more likely to be true than the null hypothesis H_0, whereas a low value of l_n indicates that the null hypothesis is more likely to be true. The testing process will continue until one of the following arises:

1. The log-likelihood ratio l_n becomes smaller than a critical minimum value A. In this case, the conclusion is that the standardized difference θ is unlikely to be as large as θ_R, and the null hypothesis H_0 will not be rejected.
2. The log-likelihood ratio l_n becomes larger than a critical maximum value B. In this case it will be concluded that there is a standardized difference between the average exposures of cases and controls as large as or larger than θ_R, and the null hypothesis is rejected in favor of the alternative hypothesis.

Whitehead (8) developed a more general approach to sequential test procedures, which includes procedures that are equivalent to Wald's sequential probability ratio tests, and which is based on a log-likelihood function (with unknown parameter θ) rather than on a log-likelihood ratio. The log-likelihood function can be expressed in terms of the parameter θ (for our comparison of two mean exposures still defined as $\theta=(\mu_1-\mu_0)/\sigma$), as well as of two test statistics, Z and V, which are both computed at each stage of the sequential test procedure. Formulae for the computation of Z and V are given in Appendix I. Z is the so-called "efficient score for θ" and, for the comparison between quantitative exposures of cases and controls discussed here, is computed as the cumulative difference in exposure divided by an estimate of the unknown standard deviation σ. V is a measure of the amount of information about θ contained in Z, also referred to as "Fisher's information", and increases as the sequential test procedure progresses. Whitehead has shown that, when θ is small and samples are large, then, at any stage in the sampling process, Z follows approximately a normal distribution with mean θV and variance V (8; pp. 60).

In practice, the sequential testing process can be conveniently presented in the form of a graph, plotting Z against V. The testing process then continues until:

1. Z becomes smaller than the critical value $A^* = -a+bV$, in which case H_0 cannot be rejected, or
2. Z becomes larger than the critical value $B^* = a+bV$, in which case H_0 will be rejected.

The critical values A^* and B^* are both linear functions of V. The slope (b) and intercepts ($\pm$ a) of these linear functions depend on the values chosen for α, β, and θ_R (see Appendix I). An example of the graphic presentation of the sequential t-test is shown in Figure 1 (further discussed in the next section). The computations for this example, including those for determination of the critical values A^* and B^*, were performed using the computer program "PEST", developed by Whitehead and Brunier (13).

An example

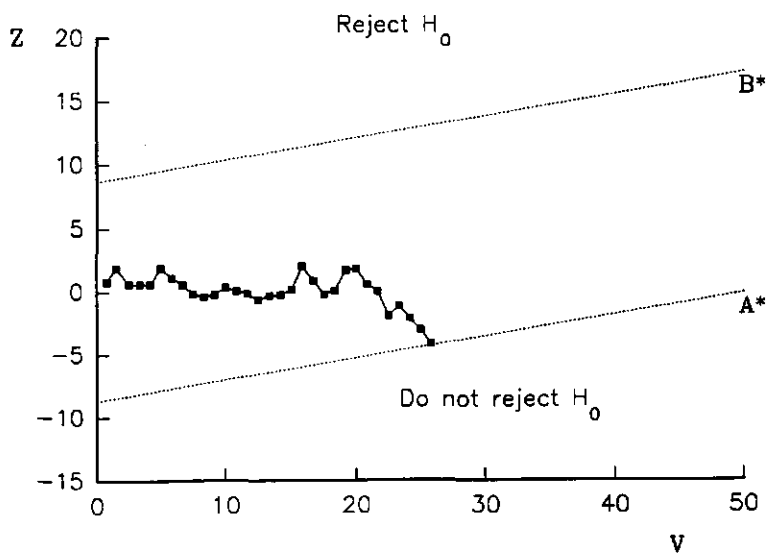
Within a cohort of participants in the "DOM"-project, a population-based breast cancer screening programme at Utrecht (the Netherlands), toenail clippings were collected and stored in a biological bank. After an average follow up of 25.7 months, a total of 61 cases of pre-menopausal breast cancer were detected.[11] Results were reanalysed using a sequential t-test. The null hypothesis of an equal selenium content in toenails of cases and of controls (H_0: $(\mu_0-\mu_1)/\sigma=0$) was tested against the one-sided alternative of a higher selenium content in the control group (H_1: $(\mu_0-\mu_1)/\sigma \geq \theta_R$). The θ_R-value was chosen equal to 0.25. The significance level and the statistical power were fixed at $\alpha=0.05$ (one-sided) and $1-\beta=0.80$, respectively. Case-control subsets consisted of one case and five controls each, and were analysed in the chronological order in which the cases had been diagnosed.

The results of the sequential testing procedure are shown in Figure 1. After a total number of 31 case-control sets (i.e., 31 cases and 155 controls), the sample path of the efficient score Z plotted against V crossed the critical boundary corresponding to no rejection of the null hypothesis.

Gain in efficiency: the expected sample size

The advantage of sequential procedures is that the expected number of observations (average sample size) needed to reject a given study hypothesis, or not, is smaller than when the test is based on a fixed sampling procedure (i.e., with predetermined sample size). Indeed, it has

Figure 1. Sample path and critical boundaries for the Selenium and Breast Cancer data (one-sided sequential t-test without matching; α=0.05, 1-β=0.8 and θ_R=0.25).



A^* and B^* are the critical boundaries of the test; Z is the so-called "efficient score" for θ, computed as the cumulative standardized difference between the exposures of cases and controls; V is a measure of theamount of information about θ contained in Z, also referred to as "Fisher's information" statistic.

been shown that, when either the null hypothesis H_0 or the alternative hypothesis H_1 is true, the sequential probability ratio test is a more efficient test (4).

Table 1 shows, for different values of θ_R, the expected sample size for the sequential t-test used in the previous example, as compared to that for a test with a fixed sample size (these expected sample size values can be computed by the PEST program).

Table 1. Expected number of case-control sets N in a sequential test for a standardized exposure difference θ, when in reality θ=0, θ=θ_R, or θ=0.75θ_R. Test without matching; α=0.05, 1-β=0.80.

		sequential test			fixed sample test
	θ_R	θ=0	θ=θ_R	θ=0.75θ_R	
one control per case (k=1)	0.15	255	361	414	550
	0.25	92	130	149	198
	0.35	47	67	76	101
five controls per case (k=5)	0.15	153	216	249	330
	0.25	55	78	90	119
	0.35	28	40	46	61

It can be seen from this table that, for a sequential t-test with the given specifications (one-sided α=0.05, and 1-β=0.80), the expected sample size under H_0 is approximately 0.46 times the fixed sample size at all values of θ_R. The expected sample size under H_1 is approximately 0.66 times the fixed sample size. For the open sequential test procedure described here, the expected sample size of the sequential t-test reaches its maximum in situations where the true θ-value is approximately equal to 0.75 θ_R, but even then remains below the sample size for a classical, fixed sample test of equal reliability.

Choice of the alternative hypothesis

In sequential test procedures, an explicit definition of the alternative hypothesis H_1 is required, specifying the minimum standardized exposure difference θ_R high enough to be detected with a given statistical power. Specification of the alternative hypothesis, in addition to the null hypothesis H_0, results in a rule which defines at which stage there is sufficient evidence for not rejecting H_0 or for rejecting H_0 in favor of H_1. If there were no such a rule for stopping a sequential test procedure without rejection of H_0, the sampling of cases and controls could continue infinitely in those situations where no difference in exposure between cases and controls exists, without ever reaching a conclusion.

The probability that, at a given stage in the sequential testing process, sufficient evidence will have accumulated on whether or not to reject the null hypothesis, H_0, depends on the specific alternative hypothesis against which H_0 is tested. For example, imagine a situation in which, at a given number of observations, there appears to be little difference between the mean exposures of cases and of controls. In such situations, the log-likelihood ratio l_n would tend to be small if the alternative hypothesis were defined by a relatively extreme θ_R-value, and H_1 would appear less likely to be true than H_0 given the case and control observations. At a small value of θ_R specified, however, the same set of case- and control-observations would have led to a higher log-likelihood ratio. The probability of concluding the test procedure with no rejection of the null hypothesis would therefore be higher in the first case (high value of θ_R) than in the second (small value of θ_R). Of course, this phenomenon is not specific for sequential tests in particular, but occurs also in statistical procedures based on a fixed sample size. The example does underline, however, that the choice of the alternative hypothesis (i.e., the value for θ_R) should be well motivated, in terms of potential public health impact or strength of the biological relation to disease.

For the sequential t-test discussed here, θ_R is specified as a standardized difference between the mean exposures of cases and of controls. For epidemiologists, who are more familiar with the definition of study hypotheses in terms of measuring disease risk, this specification of the alternative hypothesis may be difficult to interpret. However, if the disease incidence is low over the entire range of exposures (i.e., the "rare

disease" assumption), and assuming that the alternative hypothesis is true, it is possible to compute a minimum expected odds ratio value, OR_R, for different quantile levels of the distribution of exposure measurements within the cohort (from which cases and controls were drawn). For instance, the expected odds ratio for the highest versus the lowest quintile of the exposure distribution equals:

$OR_R(Q5\text{-}Q1) = e^{2.80\theta_R}$ (see Appendix II.A).

Thus, for an alternative hypothesis defined as $\theta \geq 0.25$, the minimum expected odds ratio of disease for the highest versus the lowest quintile of exposure measurements approximately equals 2.0. An extended list of expected odds ratio estimates, for different values of θ_R, is given in Table 2.

Table 2. Expected odds ratio, $OR_R[Q5\text{-}Q1]$, for the highest versus the lowest quintile of the exposure distribution in the cohort, under the alternative hypothesis $\theta=\theta_R$. Study without matching.

θ_R	$OR_R[Q5\text{-}Q1]$
0.15	1.5
0.20	1.8
0.25	2.0
0.30	2.3
0.35	2.7
0.40	3.1

Analysis of matched studies: the pairwise sequential t-test

The sequential test procedure described so far did not take account of any potential confounding factors. In many situations, however, it may be necessary to adjust for potential confounding factors such as age, duration of follow up, or additional risk factors such as body weight and menopausal status. Using the sequential procedure described here, adjustments for confounding can be made by matching cases and controls for such additional risk factors. In case-control studies nested within a cohort this may not be too complicated, since there will be a vast pool of disease-free subjects in which to find matched controls (unless there are many matching criteria). Whenever a matched study design is used, however, the matching should be

reflected in the analysis in order to obtain unbiased results.

A matched sequential t-test can be based on the pairwise differences between the exposure measurement of a case, and the exposure measurement of each of k controls belonging to the same matched subset. We shall assume that these differences, D_{ij} (where j=1,.....,k indicates the j-th control subject in the i-th case-control set; i=1,...,n), will be normally distributed:

$$D_{ij} \simeq N(\delta,\tau^2),$$

where δ is the mean, and τ^2 the variance of the differences D_{ij}. As in the unmatched situation, the hypotheses H_0 and H_1 can then be defined in terms of a standardized difference θ :

$$H_0 : \theta = \delta/\tau = 0,$$

and, for the one-sided alternative of a higher exposure for cases than for controls,

$$H_1 : \theta = \delta/\tau \geq \theta_R \ ^{(1)}.$$

The computation of the statistics Z and V is slightly different from that for the unmatched situation (see Appendix I.B). However, the formulae for the critical boundaries of the test, A^* and B^*, remain the same (since these depend only on the values chosen for α, β, and θ_R). Also, with respect to a fixed-sample test, efficiency gains will be made similar to those in the unmatched situation, in terms of a decrease in the expected sample size.

Again, with some additional assumptions it is possible to compute expected odds ratio values under the alternative hypothesis, $\theta=\theta_R$, for quantile levels of the within-stratum exposure distribution (strata being defined by the matching variables). As before, it will be assumed that for cases and controls the exposure measurements have an equal variance, σ^2, and that the overall incidence of disease is low. In addition, it will be assumed that, after matching, exposure measurements are equally correlated between controls or between cases and controls. The variance of the exposure differences D_i, between a case and k matched controls, can then be written as:

$$\tau^2 = 2\ (\sigma^2-\gamma) = 2\ \sigma'^2 \quad ,$$

where γ is the covariance between the exposure measurements of cases and controls (due to the matching), and σ'^2 is the average variance of exposure

(1) Note: If the exposure is expected to be higher for controls, the alternative hypothesis may be defined as $H_1 : \theta=\delta/\tau \leq -\theta_R$. A two-sided alternative may be specified as $H_1 : |\theta| \geq \theta_R$.

among controls (and thus, approximately, in the full cohort) within strata defined by the matching variables. The expected odds ratio for the within-stratum difference between the upper and the lower quintiles of the exposure distribution will be approximately equal to:

$$OR_R(Q5\text{-}Q1) = e^{2.80\sqrt{2}\,\theta R},$$

with $\theta_R=\delta/\tau$ (see Appendix II.B). In Table 3, some expected odds ratio values are given for different values of θ_R.

Table 3. Expected odds ratios, OR_R[Q5-Q1], for the highest versus the lowest quintile of the exposure distribution of the cohort within strata of the matching variables, under the alternative hypothesis that $\theta=\theta_R$ (1:k matching).

	OR_R[Q5-Q1]				
θ_R	k=1	k=2	k=3	k=4	k=5
0.15	1.8	1.7	1.6	1.6	1.6
0.20	2.2	2.0	1.9	1.9	1.8
0.25	2.7	2.4	2.2	2.2	2.2
0.30	3.3	2.8	2.6	2.6	2.5
0.35	4.0	3.3	3.1	3.0	2.9
0.40	4.9	3.9	3.6	3.5	3.4

Discussion

We have shown how a sequential t-test can be applied in case-control studies where the exposure measurement is a continuous variable. The use of the sequential probability ratio method in epidemiologic studies has been suggested before by O'Neill and Anello (14), who described a sequential test for analysing (matched pair) case-control studies, with a dichotomous exposure variable. So far, however, this has not been put into practice widely. An explanation may be that the advantage of a smaller expected sample size does not outweigh certain drawbacks in the use of an SPRT procedure, particularly in studies where (dichotomous) exposure assessments are based on information derived from questionnaires. One such drawback may have been the fact that epidemiologists are not familiar enough with sequential statistical methods and, until recently, no simple computer software for sequential analysis was widely available. Another drawback may

be that the sequential probability ratio procedure does not allow flexible, multivariate data modelling for the control of varying sets of confounding factors. In spite of these various drawbacks, however, a strong argument in favor of the use of sequential methods is the desire to make optimal use of material from biological banks, reducing the number of biological samples needed to test a given hypothesis.

In a sequential design, the number of case-control sets that will be sampled before a conclusion is reached is a random variable, the mean of which is smaller than the size of an equivalent fixed-sample test (as illustrated in Table 1). Occasionally, however, larger numbers of case-control sets may be needed for the test to come to a conclusion. This may introduce some uncertainty to the process of setting a budget for grant requests. However, budgets can be reasonably planned on the basis of the 90th percentiles, rather than the means, of the possible sample size distributions (assuming $\theta=0$, $\theta=\theta_R$, or $\theta=0.75\theta_R$). The PEST programme contains a sub-routine for the computation of these percentiles, at the planning stage of a study. Further details about these computations can be found in Whitehead's textbook on sequential medical trials (8).

The sequential t-tests described in this paper can be useful especially in exploratory studies, to decide, at the expense of as little biological material as possible, whether a new hypothesis seems worth further investigation, or whether it is more likely that it would eventually be proven unfruitful. It is generally agreed, however, that the use of hypothesis testing is an unsatisfactory way of assessing and presenting epidemiological findings, and that results should rather be presented as estimates of relevant measures of exposure-disease association, and their confidence intervals (15,16). Therefore, after terminating the sequential test, and irrespective of whether the null hypothesis is rejected or not, a presentation of final results should always include such point and interval estimates, describing the association between the marker values and disease risk (for instance, in terms of relative risks for different quantiles of the marker assessments). Since, on average, a sequential test will terminate at a smaller sample size than an equivalent fixed-sample procedure, estimates of epidemiological effect measures may be relatively imprecise. Once a given hypothesis has been proven of interest, however, (i.e., in case of rejection of the null hypothesis of "no difference" in exposure), the

investigator may decide to extend the number of laboratory assessments, so as to increase the precision of the study. The number of additional assessments needed to reach sufficient precision can then be determined from the standard error of effect estimates at the end of the sequential test, similarly as in a double sampling design (17).

The combination of sequential testing and subsequent estimation of epidemiologic effect measures - with or without further extension of the study - can be seen as a two-step estimation procedure, which will tend to result in effect estimates with a desired precision if there is a clear difference in exposure, or in less precise estimates if no exposure difference of interest exists. In the latter case, on average more biological samples will be saved for the investigation of other hypotheses.

O'Neill and Anello (14) have described how, for a dichotomous exposure variable and for matched case-control pairs, the critical values of a sequential test can be interpreted in terms of odds ratio values. We have shown that, under the rare disease assumption, and for a matched or an unmatched case-control design, similar interpretations can be given to the critical θ_R-value of a sequential t-test for comparison of cases and controls by a continuous exposure variable. However, due care must be taken to avoid mis-interpretation. The sequential procedures described in this paper essentially provide a test for a difference between the mean exposures of cases and of controls, and are not a substitute for a test of statistical significance for odds ratios at different quantile levels of exposure. It is possible to compute expected odds ratio values for different quantile categories of exposure, such as quartiles or quintiles, under the assumption that the alternative hypothesis $\theta=\theta_R$ is true (i.e., that a certain standardized difference in mean exposure actually exists). The relation between a θ_R-value chosen and expected odds ratio values for different quantile levels of exposure is of interest only as far as it may help define a reasonable θ_R-value for the alternative hypothesis. Within this context, the choice of quintile levels of exposure was of course quite arbitrary; computation of expected odds ratio values for tertiles or quartiles could be equally informative.

The exact value which should be chosen as a reference odds ratio value OR_R (as defined for instance for quintiles) may depend on the specific hypothesis to be tested, as well as on the potential relevance of the

exposure in terms of attributable risk (i.e., also taking into acount the prevalence of exposure within a population). O'Neill and Anello recommend specifying that the alternative hypothesis should correspond to an odds ratio not greater than about 2.0 for exposed versus non-exposed subjects (the exposure in their paper being defined as a dichotomous variable). We agree that the value of θ_R should always correspond to relatively small expected odds ratio values, so that a failure to reject the null hypothesis can be interpreted as the absence of any relevant association between exposure and disease risk. Of course it should also be kept in mind that, due to intra-individual variation over time, many biochemical markers will provide only an approximate estimate of the true risk factor of interest, and that the observed association with disease risk (also in terms of a standardized difference between mean exposures) may therefore be attenuated.

In this paper, it was assumed that the sequential testing process proceeds in steps corresponding to case-control sets consisting of only one case and its k controls. It will often be more practical, however, to run laboratory analyses in batches of more than only one case-control set at a time. It is possible to perform the sequential probability ratio test on case-control sets each comprising multiple cases. The only disadvantage of such larger inspection intervals is that there can be some "over-running" of the critical boundary, by the sample path of Z plotted against V. The number of observations may thus exceed the number that was actually required to reach a conclusion, and part of the advantage of sequential methods, in terms of a reduction in expected sample size, will be lost. However, this loss of efficiency resulting from over-running can be limited by including only a relatively small number of cases in each group of observations.

We have discussed only so-called "open" or "non-truncated" procedures, in which no upper limit has been set to the number of observations needed before a conclusion is reached. Therefore, although sequential procedures will on average require fewer case-control comparisons than equivalent tests based on a fixed sample size, there may be occasions on which the sequential procedure terminates after a much larger number of observations than would have been required for a classical, fixed sample test. In "closed", or "truncated" sequential procedures, an upper limit is fixed for the actual number of observations that may be needed in order to reach a conclusion.

For instance, it may be decided that the null hypothesis will not be rejected if the number of case-control comparisons becomes larger than twice the normal sample size for a fixed sample test without reaching the critical boundaries, A^* or B^*. Such an additional stopping rule will then affect α and β to some extent. If the maximum number of observations chosen is sufficiently large, however, the effects on these error probabilities will be relatively small. Whitehead's computer programme "PEST" (13) provides an option for the analysis of sequential studies with a truncated design. More extensive discussions of truncated sequential procedures are given in his textbook on sequential clinical trials (8), as well as by Wetherill and Glazebrook (5), and Armitage (7).

Aliquots of biological specimens such as blood serum cannot be thawed and refrozen too frequently without potentially causing changes in the biochemical parameters of interest. However, the volume of aliquots may often be sufficiently large to allow more than one type of biochemical analysis within the same laboratory. It would thus be possible to study several etiological hypotheses in parallel, based on different biochemical markers measured in the same aliquot. The simple sequential tests described in this paper are based on the concept of studying only one type of exposure measurement in relation to a single type of disease. Further development of sequential statistical methods is needed, so that such multiple, parallel hypotheses can be evaluated simultaneously with minimal loss of biological material.

References

1. de Waard F, Collette HJA, Rombach J, Baanders-van Halewijn EA, Honig C. The DOM project for the early detection of breast cancer, Utrecht the Netherlands. J Chron Dis 1984;1:1-44.
2. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). Ann Oncol 1992;3:783-791.
3. van Noord PAH. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). Thesis. University of Utrecht, The Netherlands, 1992.
4. Kendall MG, Stuart A. Sequential methods. In: Kendall MG, Stuart A. The advanced theory of statistics, Vol. 2: Inference and relationship. London: Griffin & Company, 1967:624-658.
5. Wetherill GB, Glazebrook KD. Sequential methods in statistics. London: Chapman and Hall, 1986.
6. Wald A. Sequential analysis. New York: Dover Publications Inc, 1947.

7. Armitage P. Sequential medical trials. Oxford: Blackwell Scientific Publications, 1975.
8. Whitehead J. The design and analysis of sequential clinical trials. 2nd edition. New York: Ellis Horwood Ltd, 1992.
9. Hulka BS, Wilkosky TC, Griffith JD. Biological markers in epidemiology. New York: Oxford University Press, 1990.
10. Riboli E, Rönnholm H, Saracci R. Biological markers of diet. Cancer Surveys 1987;6:685-718.
11. van Noord PAH, Collette HJA, Maas MJ, de Waard F. Selenium levels in nails of premenopausal breast cancer patients assessed prediagnostically in a cohort-nested case referent study among women screened in the DOM project. Int J Epidemiol 1987;16:318-322.
12. Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics 1974;31:643-649.
13. Whitehead J, Brunier H. PEST programme, version 2.2. University of Reading: Department of Applied Statistics, 1992.
14. O'Neill RT, Anello C. Case-control studies: a sequential approach. Am J Epidemiol 1978;108:415-424.
15. Rothman KJ. A show of confidence. New Engl J Med 1978; 299: 1362-63
16. Gardner MJ, Altman D. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986;292:746-50
17. Govindarajulu Z. Sequential statistical procedures. Chapter 4: Sequential estimation. London: Academic Press, 1975.
18. Rosner B, Hennekens CH. Analytical methods in matched pair epidemiological studies. Int J Epidemiol 1978; 7: 367-72

Appendix I. Computation of the statistics Z and V, and formulae for critical boundaries A^* and B^*.

I.A Analysis without matching: After the n-th case-control subset, the following sample statistics will be available:

	Cases	Controls	All
Number of observations	n	nk	n(k+1)
Sum of observed exposures	S_1	S_0	S
Sum of squares	Q_1	Q_0	Q

The statistics Z and V are computed from the cumulative sums, S_1 and S_0, and cumulative sums of squares, Q_1 and Q_0, of the exposure measurements of cases and controls, respectively (see Whitehead (7) pp. 57-62). The efficient score statistic Z is computed as:

$$Z = \frac{n\ k\ S_1 - n\ S_0}{n\ (k+1)\ C} \quad ,$$

where

$$C^2 = \frac{Q}{n(k+1)} - \left(\frac{S}{n(k+1)} \right)^2 \quad .$$

It is to be noted that C^2 is a maximum likelihood estimate of σ^2, under the null hypothesis $\theta=0$. Thus, Z is equivalent to the cumulative difference between the exposure measurements of cases and of controls, divided by a maximum likelihood estimate of the standard deviation σ. Fisher's information statistic V is computed as:

$$V = \frac{n\ nk}{n\ (k+1)} - \frac{Z^2}{2\ n(k+1)} \quad .$$

Coefficients for critical boundaries A^* and B^* are computed as:

$$a = \frac{\ln \left(\frac{1-\beta}{\beta} \right) + \ln \left(\frac{1-\alpha}{\alpha} \right)}{2\ \theta_R} \quad ,$$

and

$$b = \frac{\theta_R}{\ln \left(\frac{1-\beta}{\beta} \right) + \ln \left(\frac{1-\alpha}{\alpha} \right)} \ln \left(\frac{1-\alpha}{\alpha} \right) \quad .$$

For $\alpha=0.05$, $\beta=0.2$ and $\theta_R=0.25$ this leads to a=8.661 and b=0.17.

(A continuity correction is calculated as $\pm\ 0.583\ \sqrt{(V_i - V_{i-1})}$, independently of values of α, β, and θ_R; in this paper continuity corrections were of negligible magnitude, and have been ignored for the sake of simplicity.)

I.B Matched analysis: The following sample statistics can be computed after each new case-control subset :

Number of case-control sets	n
Sum of exposure differences D_i	S
Sum of squared differences D_i^2	Q

The efficient score statistic Z is computed as :

$$Z = \frac{S}{C} \quad ,$$

where

$$C^2 = \frac{Q}{n} \; .$$

Again, C^2 corresponds with a maximum likelihood estimate of the variance of the exposure differences D_i, under the null hypothesis (θ=0). Fisher's information statistic is computed as :

$$V = n - \frac{Z^2}{2n} \; .$$

(See Whitehead (8), pp. 67-68.)

Appendix II. Relation between θ_R and the expected odds ratio for the upper versus the lower quintile of exposure.

II.A Analysis without matching: Suppose that, both among cases and among controls, exposure measurements M have normal distributions with different means but with an equal variance:

$$M|\text{case} \simeq N(\mu_1, \sigma^2), \quad \text{and}$$
$$M|\text{control} \simeq N(\mu_0, \sigma^2).$$

If the probability density functions of both exposure distributions are given by $\phi_1(M) = Pr(M|\text{case})$ and $\phi_0(M) = Pr(M|\text{control})$, respectively, then, for a given difference in exposure, $\Delta = m_1 - m_0$, the odds ratio of disease can

be written as :

$$OR(\Delta) = \frac{\phi_1(m_1)/\phi_0(m_1)}{\phi_1(m_0)/\phi_0(m_0)} = e^{(\mu_1-\mu_0)(m_1-m_0)/\sigma^2} = e^{\theta\Delta/\sigma} .$$

Here, the standard deviation σ is unknown. If the disease incidence in the cohort is low, however, the distribution of exposure measurements of the controls will be approximately identical to the exposure distribution in the entire cohort. Then, for subjects belonging to different quantile categories of this distribution, the expected difference in exposure Δ can be expressed as a number of unknown standard deviations σ. The expected exposure measurement above a given cutpoint value L can be computed as the mean of a truncated normal distribution:

$$E(M|M>L) = \mu_0 + \sigma \left(\frac{\phi[(L-\mu_0)/\sigma]}{1 - \Phi[(L-\mu_0)/\sigma]} \right).$$

where $\phi[u]$ is the probability density function, and $\Phi[u]$ is the cumulative distribution function of the standard normal distribution at the point u. If L is chosen to be the cutpoint for the highest quintile of exposure, we find from the normal distribution table that $(L-\mu_0)/\sigma = 0.84$. The average exposure in the highest quintile is thus expected to be equal to:

$$E(M|M>L) = \mu_0 + \sigma \left(\frac{\phi[0.84]}{1 - 0.80} \right) = \mu_0 + 1.40\ \sigma.$$

Likewise, the average exposure in the lowest quintile is expected to be equal to $E(M|M<-L) = \mu_0 - 1.40\ \sigma$. Thus, the difference between the average exposures in the highest and lowest quintiles will be equal to $\Delta = 2.80\sigma$.

The expected odds ratio for the highest versus the lowest quintile of exposure can now be written as a function of θ:

$$OR_R[Q5\text{-}Q1] = e^{\theta\Delta/\sigma} = e^{2.80\theta}.$$

Inversely, this function can be used to compute the value for θ_R that corresponds with a minimal expected odds ratio, $OR_R[Q5\text{-}Q1]$ for the upper versus the lower quintile of exposure. For instance, an expected odds ratio of 2.00 corresponds with a standardized difference between the mean exposures of cases and controls equal to $\theta_R = \ln(2.00)/2.80 \simeq 0.25$.

II.B Matched analysis

Suppose that, in a matched pairs study, D_i is the difference between the exposure measurement for the i-th case and its matched control, and let the distribution of such differences be given by $D_i \simeq N(\delta,\tau_1^2)$. Then, as was previously derived by Rosner & Hennekens (18), the odds ratio for a difference $D_i=\Delta$ can be computed as :

$$OR(\Delta) = Pr(D_i=\Delta)/Pr(D_i=-\Delta) = \frac{e^{-\frac{1}{2}(\Delta-\delta)^2/\tau_1^2}}{e^{-\frac{1}{2}(-\Delta-\delta)^2/\tau_1^2}} = e^{2\theta_1 \Delta/\tau_1} ,$$

with $\theta_1=\delta/\tau_1$.

Assume, moreover, that for unmatched cases and controls the exposure distributions have an equal variance, and that, after matching, the exposure measurements are equally correlated between controls (if more than one control is matched per case), or between cases and controls. The variance of the exposure differences, D_i, between a case and a single matched control will then be equal to

$$\tau_1^2 = 2\ (\sigma^2-\gamma) = 2\ \sigma'^2 ,$$

where γ is the covariance between the exposures of cases and of controls (due to the matching). In this case, $\sigma'^2= \sigma^2-\gamma$ can be interpreted as the average variance of exposure measurements among controls (and thus, approximately, in the full cohort) within strata defined by the matching variables. The expected difference between the top and bottom quintiles of the within-stratum exposure distribution can then be written as

$$\Delta = 2.80\ \sigma' = 2.80\ \frac{\tau_1}{\sqrt{2}} .$$

The odds ratio corresponding with this difference in exposure equals :

$$OR_R(Q5\text{-}Q1) = e^{2\theta_1\Delta/\tau_1} = e^{2.80\ \sqrt{2}\ \theta_1}$$

with

$$\theta_1=\delta/\tau_1.$$

If k>1 controls are matched per case, the variance of the exposure differences D_i becomes smaller:

$$\tau_k^2 = (k+1)/k\ \sigma'^2 ,$$

and we can write

$$\theta_1 = \delta/\tau_1 = \sqrt{((k+1)/2k)}\ \delta/\tau_k = \sqrt{((k+1)/2k)}\ \theta_k.$$

Thus, with k controls per case,

$$OR_R(Q_5-Q_1) = e^{2.80\ \sqrt{2}\ \theta_1} = e^{2.80\ \sqrt{((k+1)/k)}\ \theta_k}.$$

Chapter 7

General discussion

General discussion

I. The assessment of dietary intake levels

Maximizing the variation in predicted intake levels

Estimating the predicted amount of variation in true intake level

Stratified sampling of calibration studies, and definition of "cohorts"

Further aspects related to validation and calibration, and topics for future research

Relative risks, and attributable fractions

Multivariate validation and calibration

Robust statistical methods for validation and calibration

II. The use of sequential study designs

III. Conclusions

Prospective cohort studies provide an ideal epidemiological approach to investigating the relation between dietary intake patterns, indicators of nutritional status, and the risk of developing chronic diseases such as cancer. To obtain sufficient numbers of cases with a specific form of disease, however, such studies must be very large (1-3), and thus require important investment for the collection of exposure assessments (the costs of follow up and statistical data analysis are much lower when passive follow-up is possible through routinely collected data). It is therefore fundamental to use an efficient study design, to optimize the amount of information obtained for a given investment of time and resources. Three main arguments can be identified, around which the efficiency aspects discussed in this thesis can be grouped:

1) approaches to maximize the amount of variation in true exposure level that is actually distinguished - or "predicted" - by exposure measurements collected at baseline; this is a way to increase the power of a cohort study (to test for the presence of diet-disease associations) while keeping its size constant;
2) approaches for the precise estimation of the distribution of predicted exposure levels; this is essential for accurate estimation of statistical power or sample size requirements for the cohort, as well as for the adjustement for biases in estimates of the (log) relative risk; and
3) the optimal balance between, on the one hand, a minimum study size to allow a minimum power for a statistical test on whether there is an association between exposure and disease risk, and, on the other hand, the number of different exposures measured.

Chapters 2 to 5 of this thesis are mainly related to the first two arguments, and discuss efficiency aspects related to the assessment of the habitual, long-term dietary intake levels of individual participants in a prospective cohort study. Chapter 6 is more related to the third argument, addressing the aspect of optimizing the number of relevant etiological hypotheses that can be evaluated when exposure assessments are based on biochemical markers measured in urine, blood, or (other) tissue specimen.

I. The assessment of dietary intake levels

Maximizing the variation in predicted dietary intake level

Two possible approaches have been discussed to maximize the amount of variation in true dietary exposure level predicted by the dietary questionnaire assessments collected at baseline.

First, one may select a dietary questionnaire method by which individuals can be ranked as precisely as possible by their true, habitual intake levels. This selection can be based on a validity study, conducted even before the cohort study is started, in which the correlation between questionnaire measurements and true intake level is estimated. In Chapter 2 it is concluded that this estimation requires a comparison of questionnaire assessments with at least two additional intake measurements, based on repeat food intake records, or on an intake record plus a biochemical marker. A crucial assumption is that the three measurements should have mutually independent random errors. As discussed in Chapter 3, it will often be difficult to conduct a validity study within a truly representative subgroup of the (planned) main study population. This may not be a major problem, however, as long as the validity study is used only to develop or select an optimal dietary questionnaire instrument, assuming that the selected method will also be the optimal one for use in the main study cohort.

A second method to increase the amount of predicted variation is to broaden the range of true dietary intake levels covered, by combining the data from multiple cohort studies conducted in populations with heterogeneous life styles and dietary habits. The analysis of such multi-cohort studies entails some specific problems, however. Stratifying the analysis by the factor "cohort" would restrict comparisons of dietary intake levels and disease outcome to those between subjects belonging to the same, restricted study population. Stratification would therefore defeat the main purpose of multi-cohort studies, which is to increase power by augmenting the range of dietary exposure levels. An alternative, "naive" approach would be to treat the data of all the cohorts combined as if they had been collected within a single study population, and to perform an analysis without stratification by cohort. This alternative approach would ignore, however, whether there is sufficient concordance between the cohort-specific

relative risk estimates for these to be combined into a single summary estimate. Likewise, no evaluation would be made of the presence of any confounding by "cohort", as a potential source of "ecological bias" (4). In Chapter 4 it is shown that the overall relative risk estimate obtained in a pooled, unstratified analysis is approximately equivalent to a weighted average of several component estimates based on

a) within-cohort variations in exposure level and disease risk of individuals, and
b) the between-cohort variation in the exposure and disease risk as measured at an aggregate level.

Only if there is sufficient concordance between the various component estimates is it valid to compute an overall, combined summary estimate of relative risk.

Estimating the predicted amount of variation in true intake level

Having taken all possible measures to maximize the amount of variation in true intake level predicted by questionnaire assessments collected at baseline (to optimize the power of tests for a diet-disease association), additional reference measurements are needed, at least in a representative sub-group to estimate the magnitude of this predicted variation. An important conclusion reached in this thesis is that, for a fixed total number of daily intake records taken as reference measurements in a calibration sub-study, the variance of predicted intake levels will be estimated most precisely when a calibration sub-study includes a maximum number of participants with only a single record each. A major additional advantage of this calibration study design is that it allows such sub-studies to be conducted more easily on a truly representative sample of cohort participants. In contrast to preliminary validity studies for the development and selection of a dietary questionnaire instrument, the representativeness of calibration studies is strictly required for the valid evaluation of study power or biases in relative risk estimates.

In multi-cohort studies, the calibration approach can be used to adjust for heterogeneity in cohort-specific relative risk estimates resulting from divergent biases due to dietary assessment errors. If the calibration is perfect, this will improve the precision of a pooled summary estimate, by a more optimal weighting of cohort-specific estimates (the weights being

inversely proportional to the variances of the predicted intake distributions, rather than to the variances of baseline dietary intake assessments). On the other hand, calibration increases the width of confidence intervals of relative risks as estimated within cohorts separately, due to imprecision in the estimation of the calibration factor.

No quantitative evaluation has been made of the potential gains in power by a more optimal weighting of cohort-specific evidence, against the losses in power that will be incurred within each cohort separately. The outcome of such evaluation will, among other things, depend on what we would assume to be the sources of heterogeneity between log relative estimates obtained in different cohorts (5,6). Assuming that true relative risks are the same in cohorts, the most precise summary estimate of (log) relative risk is obtained by weighting each cohort-specific estimate by the inverse of its variance, as described in Chapter 4. This approach is based on a so-called "fixed effect model" (6). When there is important heterogeneity between the estimates, however, even after calibration adjustments, it may be difficult to justify a single summary estimate for all cohorts combined. In this case it may be preferable to use a "random effects" model, in which both a between-cohort (extra-logistic, or extra-Poisson) variance and the within-cohort variances of (log) relative risk estimates are accounted for in deriving the weighting of the cohort-specific estimates (7,8). This aspect may require further research.

The pragmatic approach chosen in Chapter 5 for the computation of sample size requirements for calibration studies, is that the relative efficiency of calibration within cohorts should be high when the observed associations between questionnaire assessments and disease risk are relatively weak (such as for fat intake and breast cancer), because it is especially in this situation that even a modest increase in statistical power may be of interest. On the other hand, the relative efficiency may be lower in situations where the association is more significant even within a single cohort. This motivated the use of two alternative criteria to compute sample size requirements for dietary calibration studies:

a. the relative efficiency of the calibration study (which defines to what extent the precision of the adjusted estimate of log-relative risk (θ) is limited by random error in the estimation of the calibration factor), or
b. the ratio of the expected, calibrated θ-estimate divided by its

standard error (i.e., the expected t-value to test whether the calibrated θ-estimate differs from zero).

Using these criteria, the calculation of sample size requirements needs only one assumption, about the minimum level of correlation between questionnaire and reference measurements. The expected t-value is above all a criterion of statistical power more than of precision. The use of a fixed value of the expected t-value as a criterion for minimum precision of the estimated log relative risk implies that one accepts a larger margin of error when the estimate itself is larger. Preferably, precision should be defined by the absolute width of the confidence interval, and for estimates of the relative risk itself rather than for its logarithm. However, this definition would result in much more complicated calculations of sample size requirements, based on the separate specifications of an increased number of key parameters, such as the strength of the association between true dietary intake levels and disease risk, and of the expected number of cases.

Stratified sampling of calibration studies, and definition of "cohorts"

When planning the sampling scheme for a calibration sub-study, nested within a prospective study cohort, it must be anticipated that, during the analysis of the cohort study, statistical adjustments will be made for potential confounding factors such as age and sex. This has two implications for the design of calibration studies. First, the variance of a given exposure variable of interest will on average be smaller within confounder strata than in the non-stratified cohort. Consequently, the correlation between questionnaire and reference measurements adjusted for the confounding effect will tend to be weaker (9). To account for this effect, the estimation of sample size requirements for calibration studies should be based on the partial correlation between questionnaire assessments and reference measurements, adjusted for age and sex, and possibly also for other potential confounding factors. Second, the variation in true intake levels can vary across strata of main confounding factors such as age or sex or demographic sub-groups (10), whereas the magnitude of random errors may also show some variation. Thus, true intake differences may not be predicted in a uniform manner by questionnaire assessments, and, between strata, there may be different amounts of bias in relative risk estimates in the same way as this may happen between "cohorts" (which in fact can also be considered as strata, in a multi-cohort study). Perfect calibration can therefore also

result in an improved weighting of log relative risk estimates across strata defined by age, sex, or other potential confounding factors, the relative weights being proportional to the predicted intake variance in each stratum.

In an optimally designed calibration study, the relative efficiency of calibration (as defined by the ratio $\tilde{D}/D$ in Chapter 5) should be of equal magnitude across confounder strata. Assuming that the true log relative risk does not vary between strata, the relative efficiency will be constant if the numbers of subjects sampled for the calibration study are a fixed multiple of the expected numbers of cases (Chapter 5, equation 4). This underlines that, ideally, sample size requirements for calibration studies should be based on a relative efficiency criterion alone. The use of a high relative efficiency as the only criterion may however lead to excessively high sample size requirements in situations where there is a relatively strong association between baseline questionnaire assessments of intake level and disease risk (i.e., unadjusted relative risk estimates are relatively high), or where the number of cases is large. It was therefore proposed that, at the level of "cohorts", the sample size requirements for calibration studies would be truncated to a maximum level, using the expected t-value for the calibrated log relative risk estimate as an alternative criterion. This raises the question: At which sub-group level the truncation rule should be applied; that is, at what level do we wish to consider certain sub-groups to be separate "cohorts"? A guiding principle is that the true relation between diet and disease risk within cohorts is expected to be relatively homogeneous across strata of other potential confounding factors, whereas between cohorts this assumption remains to be verified. On the other hand, there are also practical considerations, such as the financial resources available. For the EPIC project, it was decided to define cohorts by country. One of the objectives of this project is to evaluate the consistency of relative risk estimates between countries, as life-style and dietary intake patterns vary considerably in different countries. An additional, more pragmatic consideration was that cohorts should be relatively independent within each country, and reach a sufficient level of power and precision at a national level. It was thus estimated that within each participating country in the EPIC project, the calibration sample will include about 4000 subjects (assuming a minimum correlation of 0.2 between questionnaire and reference measurements, this corresponds to a

relative efficiency of calibration of at least 0.90 or, alternatively, an expected t-value for the calibrated log relative risk greater than 4.0). Financial resources would have been insufficient, however, to conduct calibration studies of this size (i.e., including up to 4000 individuals) at a smaller sub-group level defined for example by regional study centre, ethnic group, or sex.

Further aspects related to validation and calibration, and topics for future research

Relative risks, and attributable fractions

A condition for using the calibration approach is that relative risks must be estimated for scaled quantitative intake differences, rather than for quantiles of the measured intake distribution. In the discussion of Chapter 3, several arguments are given as to why the first type of relative risk estimate should be preferred to that for quantiles. Nevertheless, an attractive aspect of relative risk estimates for quantiles is that these can be easily interpreted in terms of attributable fractions (11) if the quantile cutpoints are determined for the full cohort population (or for the control population in a nested case-control study). This may also explain why, in nutritional epidemiology, it has become customary to estimate relative risks for quantile categories.

For unbiased estimation of the attributable fraction from relative risks defined for quantile levels, the relative risk estimates should be adjusted for attenuation bias, which, for this form of relative risk estimate, requires an estimate of the correlation between questionnaire assessments and true intake values (12,13). This remains a valid reason for conducting a dietary validity study within a cohort, based on at least two reference measurements per person (e.g., daily intake records), or combining one reference measurement with a biochemical marker, as described in Chapter 2. In this context, it may be of interest to note that a validity study is also needed for the unbiased estimation of the attributable fraction for subjects with true intake levels above or below a given absolute cutpoint value T_0. Under the model assumptions of Chapters 3 to 5 – i.e., an exponential risk model, and a normally distributed intake variable – the attributable fraction F can be computed as (14)

$$F=1-\exp[\theta(T_0-\mu_T-\tfrac{1}{2}\,\theta\sigma_T^2)] \ .$$

A calibration study allows an unbiased estimation of only two of the three unknown parameters in this equation: the log relative risk θ, and the mean true intake value μ_T. A validity study will be required, however, to estimate the variance of the true intake distribution σ_T^2 which, together with the mean intake value μ_T, defines the proportions of individuals with true intake levels above or below the cutpoint value T_0.

Multivariate validation and calibration

"Validation" is usually defined as the evaluation of whether a given method actually measures what it purports to measure (15,16). In practice, one of the objectives of conducting a dietary validity study is to estimate the amounts of "noise" and "signal" in measured intake levels of foods or nutrients; that is, to separate variation due to error, from variation due to true between-subject intake differences. Throughout this thesis, validation has been considered only in terms of a univariate measurement error model, considering the intake level of only one food or nutrient. This univariate approach does not address the question whether the "signal" represents differences specifically in the type of intake variable that one purports to measure. For example, there can be high correlations between intake measurements of animal protein and saturated fat, or between vitamin C and beta-carotene, even when the two variables are measured by different methods. Partly, this correlation may be explained by the fact that, depending on body size and physical activity, some individuals consume more food than others, and that generally the intake levels of most nutrients are positively correlated with total energy intake (17). An additional explanation is that specific nutrients tend to be found in similar types of food. For example, fruits and vegetables are by far the main sources of vitamin C and beta-carotene, while meat or dairy products provide only negligible amounts of these compounds. More research is needed on the use of validity studies to estimate how much variation exists in the true intake level of one nutrient independently from that of another, using a multivariate measurement error model similar to that in Chapter 2, but with intake levels of different nutrients represented by multiple (correlated) latent variables.

The presence of multiple correlations between the many different chemical constituents of foods also form a problem for the calibration approach. For example, questionnaire assessments of animal protein intake will not only

predict true intake differences for animal protein itself, as a main factor of interest, but also for saturated fat or other constituents that may be particularly abundant in meat, eggs, or dairy products. To account for the multivariate correlations between the intakes of different nutrients as potential predictors of disease risk, these variables can be treated as mutually confounding factors, by including them simultaneously in a relative risk estimating model. A complication, however, is that each intake variable will be measured with substantial amounts of random error. When errors are independent, inclusion of one variable as a potential confounder of the effect of another will result in only a partial adjustment, leaving residual confounding (9,18-20). The situation becomes even more complex when one considers that errors in questionnaire assessments are likely to be correlated for nutrients that tend to be present in the same types of food (19-21). As a possible solution to this problem, Rosner et al. (22) have extended the linear approximation approach to the situation with multiple, correlated exposure factors each measured with error. This multi-variate calibration approach estimates the variation in the intake levels of multiple nutrients (measured by a reference method) as predicted by a similar number of baseline questionnaire assessments, and provides valid, mutually adjusted relative risk estimates. The major requirement remains that, for each nutrient, errors in the reference measurents must be independent of those of the baseline questionnaire assessments. More research is needed for the evaluation of sample size requirements for calibration studies with multiple covariates.

Robust statistical methods for validation and calibration

The approach for validation of dietary questionnaire assessments, as described in terms of structural equation models in Chapter 2, depends on the assumptions that the relations between different types of measurements are linear, and that distributions of the latent, true intake variable as well as of measured intake values are approximately normal. These have been the underlying assumptions for the analyses of most validity studies published so far, although this has not always been made very explicit. Similar assumptions are needed for the calibration approach described in Chapter 3.

In practice, the assumptions of normal distributions and of measurements having homoscedastic random errors do not always appear to be valid.

Distributions of nutrient intake assessments often show a negative skewness, reflecting a larger variance of random errors at the higher intake levels. In Chapter 2, transformations were used to improve the normality of the measured intake distributions. It is unclear, however, whether the assumption of linear relations between different types of intake measurement, or between intake levels and the logarithm of disease risk, can reasonably be made after such transformations. Future work should explore the use of more robust statistical methods for validation and calibration of dietary intake assessments, which depend less on assumptions of normality the intake distribution, and homoscedasticity of measurement errors.

II. The use of sequential study designs

An important approach to reducing the cost of a prospective cohort study is to bank "raw material" collected at baseline and to complete the exposure assessment when it is known which individuals have developed a specific form of disease, and who are suitable control subjects. This approach may in principle apply to all types of information obtained, whether collected by questionnaires or by means of biochemical markers. Coding and entry of questionnaire data into the computer may be too expensive to complete for all participants in a prospective cohort study. It may therefore be decided to complete coding and data entry only for cases with disease, as soon as these have been identified, and for a subset of disease-free subjects used as controls in a nested case-control, or case-cohort design (3). For similar reasons – in particular, the high cost of laboratory analyses – biological specimens such as blood or urine may be frozen and stored in a biological bank until it is known who has developed a specific form of disease and who are suitable control subjects. An additional reason for creating a biological bank, however, is that only a limited amount of biological specimens (e.g., blood, or urine) can be taken from each individual. To some extent, an analogous problem exists when using questionnaires for the collection of exposure information (about diet, as well as about many potential confounding factors), since including too many questions may reduce the quality of response, or may decrease rates of participation in the study. Nevertheless, an obvious difference between questionnaires and biological specimens is that a choice of questions to be included in a questionnaire must be made at the beginning of the study, whereas it can be decided later what types of biochemical markers will be assessed, once cases

and controls have been identified. Thus, the banking of biological material is used not only to reduce the costs of exposure assessment, restricting the assessments to cases and a subset of disease-free individuals, but also to postpone the decision on which hypothesis will be tested (and which corresponding markers of exposure will be assessed) depending on the type of disease outcome observed.

In the case of a well established biological hypothesis, which has long been waiting for a more definite answer (e.g., free estrogens and breast cancer risk), a precise estimation of the association between the marker and disease risk is of interest not only when this association is clearly present, but also when, after careful evaluation, the association appears to be very weak or even absent. In the case of a more tentative hypothesis, however, related to a new type of marker, one will generally be more interested in such precise estimation when a clear association does exist, whereas in the absence of a clear association one would rather save the biological samples to search for stronger predictors of disease risk. The sequential t-test discussed in Chapter 6 presents a simple approach to deciding whether a new hypothesis is worth further investigation, while avoiding wasting too much biological material in testing a hypothesis which is not strongly supported by the emprical data at hand.

The sequential probability ratio procedure presented in Chapter 6 has the advantage that on average it requires fewer observations to test for the presence of an association than a traditional, fixed sample test procedure, not only in situations where the null hypothesis is to be rejected, but also in situations where the null hypothesis is true. The latter is not true of all sequential procedures, however. For example, Pasternak and Shore (23) have proposed the use of repeat significance tests on the accumulating data in prospective studies, with adjustment of the nominal significance (i.e., α-) levels for planned interim tests, to avoid an increase in the overall probability of falsely rejecting the null hypothesis (24). The appeal of Pasternack and Shore's sequential procedure is that it uses standard statistical test methods common in epidemiology. However, although the expected numbers of observation required in repeat significance tests are smaller than in fixed-sample test procedures when the alternative hypothesis is true (and the null hypothesis is to be rejected), the average number of observations needed is actually larger when the null hypothesis is true.

Chapter 6 is an extension of a previous paper which discusses two different versions of a sequential t-test, based on alternative approximations of the log likelihood ratio, and using a self-written computer programme. A copy of this paper is included in the Annex of this thesis. Computer simulations were carried out to evaluate the operating characteristics of the two alternative tests, in terms of their true levels of statistical significance and power. Similar simulations have been done to evaluate the sequential t-test by Whitehead's approach discussed in Chapter 6 (i.e., using the "PEST" programme), and it was found that generally this approach is superior to the methods used in the previous paper (van der Tweel, personal communication). An additional aspect, which had not been addressed in the previous paper, is to find a reasonable definition of the alternative hypothesis expressed as a standardized exposure difference. In Chapter 6 it is shown that this standardized difference can be related to an expected odds ratio of disease, for quantile categories of the exposure distribution.

As already mentioned in Chapter 6, a potential shortcoming of the proposed sequential t-test is that it allows an evaluation of only one type of exposure at a time, whereas in practice it is often possible to measure several markers in the same aliquot of a biological specimen. More work is therefore needed on the use of sequential methods in which the stopping rule is based on case-control differences in more than one type of exposure.

III. Conclusions

A basic approach to improving the statistical power of a cohort study without increasing its size is to maximize the amount of variation in true intake level predicted by measurements collected at baseline. Preliminary validity studies, based on multiple, additional measurements with independent sources of error, can help select an optimal questionnaire instrument to measure dietary intake. Additional, unbiased ("reference") measurements are also needed to evaluate the statistical power and sample size requirements of a cohort study, and to obtain unbiased relative risk estimates. For the latter two objectives, however, it is more efficient to conduct "calibration" sub-studies based on only a single reference measurement per subject (but on a larger number of individuals). Calibration studies should always be conducted on a representative sub-sample of cohort participants. In multi-cohort studies, calibration of intake assessments can

help decrease between-study heterogeneity in relative risk estimates due to bias, and can thus improve the precision of a pooled summary estimate. Sample size requirements of calibration sub-studies can be determined on the basis of a trade-off between relative efficiency criterion or, alternatively, a minimum absolute level of statistical power for a test on diet-disease association after calibration. For optimal efficiency, the number of participants in calibration sub-studies within cohorts should be proportional to the numbers of cases expected within strata of main confounding factors.

An important aspect of the planning of prospective studies is to find an optimal balance between the cohort size required to attain a minimum level of power and precision, and the number of different exposures measured. When exposure measurements are based on the chemical analysis of biological specimens, stored in a biological bank, a sequential statistical design can be used to minimize the average number of specimens required for the preliminary evaluation of a scientific hypothesis. Thus, a maximum number of scientific hypotheses can be addressed with a given total amount of biological material available. A commercially available computer programme, "PEST", can be used for the analysis of such sequential studies.

References

1. Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Hennekens CH, Speizer FE. Dietary fat and the risk of breast cancer. N Engl J Med 1987; 316:22-8.
2. Howe GR, Friedenreich CM, Jain M, Miller AB. A cohort study of fat intake and risk of breast cancer. J Natl Cancer Inst 1991; 83:336-40.
3. van den Brandt PA, Goldbohm RA, van 't Veer P, Volovics A, Hermus RJJ, Sturmans F. A large-scale prospective cohort study on diet and cancer in the Netherlands. J Clin Epidemiol 1990; 43:285-95.
4. Piantadosi S, Byar DP, Green SW. The ecological fallacy. Am J Epidemiol 1988; 127:893-904.
5. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987; 9:1-30.
6. Spector T, Thompson SG. The potential and limitations of meta-analysis. J Epidemiol Comm Health 1991; 45:89-92.
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clin Trials 1986; 7:177-88.
8. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. Stat Med 1989; 8:141-51.
9. Liu K. Measurement error and its impact on partial correlation and multiple linear regression analysis. Am J Epidemiol 1988; 127:864-74.

10. Wassertheil-Smoller S, Davis BR, Breuer B, Chang CJ, Oberman A, Blaufox MD. Differences in precision of dietary estimates amonh different population subgroups. Ann Epidemiol 1993; 3:619-28.
11. Breslow NE, Day NE. Statistical methods in cancer research. Volume I. The analysis of case-control studies. IARC Scientific Publications No. 32. Lyon: International Agency for Research on Cancer, 1980.
12. Walker A, Blettner M. Comparing imperfect measures of exposure. Am J Epidemiol 1985; 121:783-90.
13. de Klerk NH, English DR, Armstrong BK. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. Int J Epidemiol 1989; 18:705-12.
14. McKeown-Eyssen GE, Thomas DC. Sample size determination in case-control studies: the influence of the distribution of exposure. J Chron Dis 1985; 38:559-68.
15. Block JA., A review of validation of dietary assessment methods. Am J Epidemiol 1982; 115:492-505.
16. Cameron ME, van Staveren WA (eds). Manual on methodology for food consumption studies. Oxford: Oxford University Pres, 1988.
17. Willett WC, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986; 124:17-27.
18. Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol 1980; 112:564-9.
19. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. Stat Med 1989; 8:1151-65.
20. Armstrong BG. The effects of measurement errors on relative risk regressions. Am J Epidemiol 1990; 132:1176-84.
21. Phillips AN, Smith GD. Bias in relative odds estimation owing to imprecise measurement of correlated exposures. Stat Med 1992; 11:953-61.
22. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990; 132:734-45.
23. Pasternack BS, Shore RE. Group sequential methods for cohort and case-control studies. J Chron Dis 1980; 33:365-73.
24. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J R Stat Soc A 1969; 132:235-44.

Annex

Application of a sequential t-test in a cohort-nested case-control study with multiple controls per case.

J Clin Epidemiol Vol. 46, No. 3, pp. 253–259, 1993
Printed in Great Britain.

0895-4356/93 $6.00 + 0.00

APPLICATION OF A SEQUENTIAL *t*-TEST IN A COHORT NESTED CASE–CONTROL STUDY WITH MULTIPLE CONTROLS PER CASE

INGEBORG VAN DER TWEEL,[1] PAUL A. H. VAN NOORD[2] and RUDOLF KAAKS[3]

[1]Centre for Biostatistics, University of Utrecht, Centrumgebouw Noord C122, Padualaan 14, 3584 CH Utrecht. [2]Department of Epidemiology, University of Utrecht, Preventicon, Radboudkwartier 261-263, 3511 CK Utrecht, The Netherlands and [3]IARC/WHO, Unit of Analytic Epidemiology, 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France

(Received in revised form 19 August 1992)

Abstract—Application of sequential analysis may avoid unnecessary experimentation and achieve economical use of available biomaterial stored in biological banks. When, as often happens in cohort case–control studies, cases are scarce, it may be possible to use multiple control observations per case to increase the power of a test for detecting differences between cases and controls. Samples from a biological data bank were analysed. We compared results of a non-sequential analysis with results of sequential *t*-tests for 1 to 5 controls matched per case in a cohort nested case–control study. Simulations are performed to get an idea of the unreliability and the power of the sequential test. In general the sequential *t*-tests are too conservative with respect to the achieved power. Average sample numbers are lower for the sequential tests and decrease with multiple controls. More than 3 or 4 controls per case does not give a meaningful increase in efficiency.

Sequential *t*-test Multiple controls Simulations Efficiency Biobanking Cohort nested studies

INTRODUCTION

Sequential analysis of quantitative data has never found wide application in clinical trial practice, even though considering its use might be worthwhile. For ethical reasons alone one may wish to minimize the expected number of exposed patients. From an experimental point of view, one may wish to avoid unnecessary experimentation. In cohort nested case–control studies exposures may be assessed in biological samples stored in a biological bank. In this situation, economy with material from the biological bank may be a reason to choose a sequential type of analysis. In a prospective study, cases are often detected sequentially during follow-up. A sequential analysis could then limit the total duration of the study.

In a sequential case–control analysis, the response of a case is compared with the response of a single control. O'Neill describes in a detailed way a sequential analysis of a matched pair case–control study with a dichotomous response [1].

In a cohort study, usually there is only a limited amount of biological material per subject, and there are far more controls in the biobank for which such material can be analyzed than cases. Therefore it may be desirable to compensate for the loss of statistical power by comparing each case with more than one control [2].

Ury [3] showed that, for non-sequential case–control studies with continuously distributed data, the efficiency of multiple ($k > 1$) controls relative to matched pairs ($k = 1$) is equal to $2k/(k + 1)$ for equal case and control variability.

Gail *et al.* [4] show that in (non-sequential) situations with a limited number of cases, more than four controls per case (or vice versa) does not give a meaningful power increase.

We are unaware of literature about the efficiency of multiple controls per case in sequential analyses. Therefore, we compared the effect of more controls per case in a sequential design with the results of a non-sequential analysis.

MATERIALS AND PATIENTS

We performed retrospective analyses on data from a cohort nested case–referent (control) study on breast cancer and the selenium content in ppm of toenails (Van Noord [5]). The aim of the study was to determine whether selenium, as available in the body, is already decreased before tumour occurrence.

Nail clippings had been collected since 1982 in a cohort of 8760 premenopausal (i.e. without menopausal signs) women (42–52 years of age), who attended a breast cancer screening program. A total number of 64 premenopausal breast cancer cases were detected in this cohort. Controls were matched to cases for age. For 57 cases 5 controls per case were available; for 7 cases 3 or 4 controls could be matched per case.

Selenium content in the nails did not depend on age, probably due to the relatively small age-range in our data. No seasonal or other time trends were found in nail selenium contents during 3 years of investigation (unpublished results).

The data were analysed in the order the cases became available over time.

STATISTICAL ANALYSIS

Non-sequential analysis

For matched case–control observations the minimal sample size n_1 (i.e. the number of case–control pairs necessary) for detecting a true difference between case and control observations of at least μ with a (two-sided) type I probability (or unreliability) α and a type II probability β (i.e. power $1-\beta$) is [6]

$$n_1 = (t_\alpha + t_\beta)^2 * \sigma_1^2/\mu^2,$$

where

σ_1^2 is the variance of the difference between a case and a control observation,

t_α and t_β are values from the two-tailed t-table with $n_1 - 1$ *df* corresponding to probabilities of α and β respectively.

The type I probability α is the risk one wants to accept that the null hypothesis of no difference between case and control observations is falsely rejected; the type II probability β is the risk of falsely not rejecting the null hypothesis when a true difference of at least μ exists between case and control observations.

In case of multiple (say k) control observations per case, assuming equal variances for cases and controls and, for the sake of argument, a negligible correlation between case and control observations, the variance of the difference between a case observation and the mean of the k control observations becomes

$$\sigma_k^2 = \{(k+1)/k\} * \sigma^2 = \{(k+1)/2k\} * \sigma_1^2,$$

($\sigma_1^2 = 2\sigma^2$, where σ^2 is the variance of a single case or control observation).

The minimal number of case–control sets for detecting the same difference μ then becomes

$$n_k = (t_\alpha + t_\beta)^2 * \sigma_k^2/\mu^2 = n_1 * \{(k+1)/2k\}.$$

N.B. We assumed (near) independence of case and control observations. In case of a positive correlation between case and control observations, the result will be a smaller σ_1^2 and σ_k^2 and a smaller sample size needed to detect the same difference μ.

Sequential analysis

Wald [7] developed the theory for the "sequential probability ratio test" (SPRT). Rushton [8] further developed this theory to the one-sample, two-sided sequential t-test. This test is based on the probability ratio

$$l_n = \frac{\text{probability of observed results given } H_1 \text{ true}}{\text{probability of observed results given } H_0 \text{ true}},$$

for n observations processed so far. For our situation with case–control sets, we pose as null hypothesis H_0:

$$\delta = \mu/\sigma_k = 0$$

and as alternative hypothesis H_1:

$$|\delta| > 0$$

where μ is the minimal mean difference to be detected and σ_k is the theoretical standard deviation of the differences between the case and control observations. Because in most practical

situations σ_k will be unknown and needs to be estimated from the data, the parameter $\delta = \mu/\sigma_k$ is used in the test. The test operates as follows:

—continue sampling as long as $B < l_n < A$
—stop sampling and decide for H_0 as soon as $l_n < B$
—stop sampling and decide for H_1 as soon as $l_n > A$

To obtain approximately the *a priori* specified error probabilities α (two-sided type I error) and β (type II error), Wald stated the theorem that $A \simeq (1-\beta)/\alpha$ and $B \simeq \beta/(1-\alpha)$. The logarithm of the probability or likelihood ratio l_n can be calculated exactly using the series expansion of Kummer's function [9].

Rushton [8] obtained a practical approximation to the logarithm of the likelihood ratio.

See Appendix A for more details on Kummer's function, Rushton's approximation and our adaptation of the test statistic for k control observations per case.

Simulations

To examine the effect of multiple controls per case in a sequential t-test on its overall type I and type II error, simulation studies were performed. A simulation program was written in Turbo Pascal Version 5.0 (Borland). Random case and control observations were generated following a normal distribution with expectation μ_0 or μ_1 and theoretical standard deviation σ. The values chosen for μ_0, μ_1, σ and δ under H_1 are based on population values and a desirable shift in ppm of the selenium content (see Van Noord [10]). Both for case and control observations σ was chosen equal to 0.15. Under H_0: $\delta = 0$, μ_0 was chosen equal to 0.8. Under H_1: $|\delta| = \delta$, μ_1 was equal to $0.8 + \delta * \sigma * \sqrt{2}$.

Both under H_0: $\delta = 0$ and under H_1: $|\delta| = \delta$ ($\delta = 0.3$, 0.4 and 0.5 respectively), and with 1 to 5 controls per case, we ran 1000 simulation runs ($\alpha = 0.05$, $1 - \beta = 0.80$).

Per run, the resulting decision ("accept H_0" or "reject H_0 in favour of H_1") and the number of case–control sets necessary to come to that decision were recorded.

Simulations were performed using both Rushton's approximation to the logarithm of the likelihood ratio and the series expansion of Kummer's function.

RESULTS

Non-sequential analysis

The results of a randomized block analysis of variance on the "selenium and breast cancer" data for $n = 57$ cases and 5 control observations per case are shown in Table 1.

Table 1. "Selenium and breast cancer" study; descriptive statistics and ANOVA table for 57 cases with 5 controls per case

	mean (ppm)	SD (ppm)	n		
Cases	0.790	0.156	57		
Controls	0.772	0.207	285		
	ANOVA table				
Source	Sum of squares	Degrees of freedom	Mean squares	F	p
Between matched sets	2.20	56	0.04		
Within matched sets	0.16	5	0.03	<1	NS
Case–controls*	0.02	1	0.02		
Between controls	0.14	4	0.03		
Residual	11.24	280	0.04		

*Due to the difference between cases and the mean of the matched control observations.

Means, standard deviations and a randomized-block analysis of variance (ANOVA) table for $n = 57$ cases with 5 controls per case. Data are the selenium content in ppm in toenails from the "selenium and breast cancer" study.

Within matched sets the sum of squares, degrees of freedom and mean square are subdivided into two components: one that measures the variation because of a difference between cases and the mean of the matched control observations, and one that measures variation between controls. If we assume no differences between control observations, this last component can be combined with the residual sum of squares to give a (slightly) improved estimate of the residual mean square or error variance.

The mean difference between a case and the mean of the corresponding 5 control observations was 0.018 ppm with a SE = 0.029 ppm (NS).

Sequential analysis

Sequential t-tests were performed on the "selenium and breast cancer" data, using the available cases and a random sample of k ($k = 1, \ldots 5$) control observations available in the matched set. (For each sequential test performed, control observations were replaced.) Both Kummer's function and Rushton's approximation were applied.

The number of cases (n) at which the decision "H_0 cannot be rejected" was reached, is tabulated in Table 2 for several alternative hypotheses ($|\delta| = 0.3, 0.4, 0.5$).

N.B. None of the tests led to rejection of H_0; in the case of $H_1: |\delta| = 0.3$, for some tests no conclusion could be reached with the available number of case–control sets.

Simulations

The relative efficiency of more (k) controls per case is depicted graphically in Figs 1 and 2 for $\delta = 0.4$. (For $\delta = 0.3$ and $\delta = 0.5$ the course of the relative efficiency is similar.) There the relative sample size n_k/n_1 is plotted against k for the median, mean and 95th-percentile number of cases required to reject H_0 in favour of H_1. The theoretical expected efficiency $(k + 1)/2k$ is plotted as a comparison.

Appendix B shows data and calculations of one of the simulations as an example.

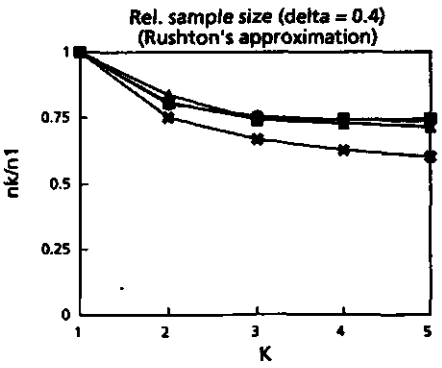


Fig. 1. Relative sample size (n_k/n_1) for mean (▲), median (●) and 95th percentile (■) number of cases necessary to reject H_0 in favour of $H_1: |\delta| = 0.4$ compared to the theoretical expected value $(k + 1)/2k$ (✖), using Rushton's approximation.

DISCUSSION

Biological data banks contain valuable material that can be analysed to explore new hypotheses with possible important public health consequences. But, with most chemical analyses, these unique biological samples are destroyed and thus economical tests are preferable [11].

While in case–control studies, cases are mostly scarce, but control samples abundant, statistical efficiency of non-sequential tests can be increased by including multiple controls per case. If the power using equal allocation ($k = 1$) is greater than 0.9, this is of no practical importance. If the equal allocation power is less than 0.9, meaningful power increases may be obtained, but more than 4 controls per case are seldom worthwhile [4].

Table 2. "Selenium and breast cancer" study; results of sequential t-tests for k controls per case

	H_1					
	$\|\delta\| = 0.3$		$\|\delta\| = 0.4$		$\|\delta\| = 0.5$	
k	R	K	R	K	R	K
1	21	—	12	23	9	13
1	25	—	30	30	11	15
1	27	62	21	21	10	13
1	22	48	23	24	10	14
1	26	50	13	25	8	18
2	25	50	17	21	9	14
3	22	—	18	21	12	16
4	22	—	12	21	8	13
5	22	—	13	21	9	13

Results of the sequential t-tests, given 57–64 cases and random samples of k controls per case, on the "selenium and breast cancer" study ($\alpha = 0.05$ and $1 - \beta = 0.80$); R, Rushton's approximation; K. Kummer's function.

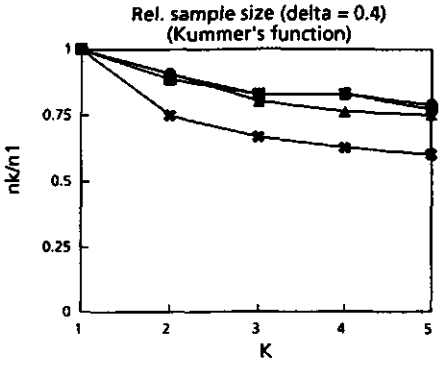


Fig. 2. Relative sample size (n_k/n_1) for mean (▲), median (●) and 95th percentile (■) number of cases necessary to reject H_0 in favour of $H_1: |\delta| = 0.4$ compared to the theoretical expected value $(k + 1)/2k$ (✖), using Kummer's function.

Retrospective analyses as well as prospective studies justify the use of sequential investigation to avoid unnecessary destruction of the biological material and to limit the total duration of the study. In prospective clinical trials ethical aspects may play a role. For example when chemotherapy is one of the trial arms in a trial comparing two cancer therapies, one wishes to expose as few patients as necessary in coming to a decision.

From an economical point of view we performed sequential t-tests with multiple control observations per case and compared the results with those of a non-sequential analysis and of simulation studies.

The expected average sample numbers (ASN) for a sequential t-test with one control observation per case are already smaller than the minimal sample size required for a corresponding non-sequential (=fixed sample size) paired t-test (Table 3). (See Appendix C for the calculation of the ASN according to Cox' approximation [12].) Notable in Table 3 is the fact that both the mean number of case–control pairs required to reject H_0 using Rushton's approximation and the median number using Kummer's function almost equal Cox' approximated ASN. Only the median number of cases necessary to accept H_0 using Rushton's approximation resembles the corresponding ASN according to Cox. Our simulations indicate that Cox' approximation probably underestimates the average sample size, especially the expected ASN needed to accept H_0.

Table 3. Comparison of expected and observed sample size for one control matched per case ($k = 1$)

	H_1		
	$\lvert\delta\rvert = 0.3$	$\lvert\delta\rvert = 0.4$	$\lvert\delta\rvert = 0.5$
Fixed			
Paired t-test	88	50	32
Sequential			
Expected:			
Cox' approximation	57/34	34/20	22/14
Observed:			
Simulation results			
Rushton			
Mean	57/44	36/27	25/18
Median	50/33	31/20	21/13
Kummer			
Mean	64/54	39/31	26/21
Median	57/43	35/25	23/17

Sequential sample sizes are expressed as "number of case–control pairs neccesary to reject H_0/number of case–control pairs neccesary to accept H_0".

Expected sample size for a non-sequential paired t-test and expected and observed sample sizes for sequential t-tests with matched pairs (i.e. 1 control per case).

Most sequential t-tests of our "selenium and breast cancer" data (Table 2) resulted in acceptance of H_0 at a considerably smaller number of case–control sets than necessary for a non-sequential analysis.

The simulations confirm these results even better. The largest gain in efficiency as compared to matched pairs is reached with 2 controls per case, when H_0 is rejected. When H_0 cannot be rejected, the gain in efficiency is smaller. The simulated power values are closer to each other for different values of δ using the exact Kummer function than they are using Rushton's approximation.

Rushton's approximation, on the other hand, is less conservative with respect to the simulated power and thus more economical in its use of case–control sets. Only with the matched-pairs simulations Rushton's approximation yields a simulated power significantly less than the theoretical power of 0.80. In general, the simulated unreliability using Rushton's approximation is larger than that using Kummer's function and more often even larger than the theoretical unreliability of 0.05.

Skovlund and Walløe [13] already drew attention to the conservatism of the sequential t-test when applied as a two-sample sequential test. Their smallest value for δ studied was 0.5, however. Neither did they simulate with more than 1 control matched per case.

In theory it is possible that a sequential test continues infinitely. To warrant that a decision is reached, albeit "no decision can be made", it is recommended to set a restriction (e.g. once or twice the fixed sample size) to the total number of cases available for the test.

Our simulations illustrate that there is hardly any effect on the simulated power and unreliability when the sequential test procedure is truncated at twice the fixed sample size.

Truncating the procedure at a fixed sample size results in a simulated power that is still too large, except for the matched-pairs situation using Rushton's approximation where it is too small. The unreliability resulting from the simulations using Rushton's approximation with more than one control per case is often (significantly) too large.

When a sequential test is terminated after a small number of observations, point and interval estimates of the case–control difference are rather imprecise. We hold the view that these objections play a less important role when, as in our experimental set-up, a rather "qualitative"

answer ("H_0 can/cannot be rejected") suffices to distinguish promising new hypotheses from unfruitful ones (see for an example Van Noord [10]).

Group sequential procedures (for matched case–control sets) [15–18] also have the advantage of a reduction in the average sample size as compared to fixed-sample-size plans. There are some differences between group sequential procedures and a one-at-a-time SPRT, however. A one-at-a-time sequential approach can be stopped after every new case–control set, while a group sequential procedure can only be stopped after the next planned inspection. Furthermore, a group sequential procedure cannot come to a decision to accept the null hypothesis until after the last planned inspection. A SPRT can be stopped the very moment that evidence exists that the null hypothesis cannot be rejected anymore.

Therefore, the authors prefer a one-at-a-time SPRT over the group sequential procedure when ethical and/or economical motives play a role. Promising hypotheses as well as unfruitful ones can be distinguished with as little as possible biological material destroyed or, for that matter, time and/or money spent.

Following Skovlund and Walløe [14], we hold the view that a sequential design might be considered more often in prospective clinical trials as well as in (cohort-nested) case–control studies.

Furthermore, we are of opinion that a sequential t-test with 2–4 controls per case is appropriate in case–control studies and other experimental designs where the case material must be used economically, and the response is available (almost) immediately. In general the investigation can then be stopped at a lower average sample size as compared to one control per case or a non-sequential test.

The use of exact calculations (the series expansion of Kummer's function) is recommended, although less conservative procedures are to be developed.

Tables and figures summarizing the results from the computer simulations are available from the authors by written request.

CONCLUSIONS

(1) A sequential t-test with 2–4 controls matched per case in general leads to lower average sample sizes than a matched-pairs sequential t-test or a non-sequential analysis. The largest gain in efficiency as compared to matched pairs is reached with 2 controls per case.
(2) Rushton's approximation to the logarithm of the likelihood ratio is rather inaccurate and leads to a power that is significantly too small in case of a matched-pairs analysis.
(3) The use of Kummer's function (the exact calculation) results in power values which are too conservative.
(4) Cox' approximation to the expected average sample number probably underestimates the expected sample size needed to accept H_0.

Acknowledgement—The study was supported by the Praeventiefonds (The Netherlands), grant No. 28-1560.

REFERENCES

1. O'Neill RT, Anello C. Case–control studies: a sequential approach. **Am J Epidemiol** 1978; 108: 415–424.
2. Lachin J. Introduction to sample size determination and power analysis for clinical trials. **Contr Clin Trials** 1981; 2: 93–113.
3. Ury HK. Efficiency of case–control studies with multiple controls per case: continuous or dichotomous data. **Biometrics** 1975; 31: 643–649.
4. Gail M, Williams R, Byar DP, Brown C. How many controls? **J Chron Dis** 1976; 29: 723–731.
5. Noord PAH van, Collette HJA, Maas MJ, Waard F de. Selenium levels in nails of premenopausal breast cancer patients assessed prediagnostically in a cohort-nested case–referent study among women screened in the DOM project. **Int J Epidemiol** 1987; 16: 318–322.
6. Sokal RR, Rohlf FJ. **Biometry**, 2nd edn. New York: W. H. Freeman; 1981.
7. Wald A. **Sequential Analysis.** New York: John Wiley; 1947.
8. Rushton S. On a two-sided sequential t-test. **Biometrika** 1952; 39: 302–308.
9. Abramowitz M, Stegun IA. **Handbook of Mathematical Functions.** New York: Dover Publications; 1968.
10. Noord PAH van. Selenium and human cancer risk (nail keratin as a tool in metabolic epidemiology). Thesis. The Netherlands: University of Utrecht; 1992.
11. Wald NJ. Use of biological sample banks in epidemiological studies. **Maturitas** 1985; 7: 59–67.
12. Wetherill GB, Glazebrook KD. **Sequential Methods in Statistics,** 3rd edn. London: Chapman and Hall; 1986.
13. Skovlund E, Walløe L. A simulation study of a sequential t-test developed by Armitage. **Scand J Stat** 1987; 14: 347–352.
14. Skovlund E, Walløe L. Sequential or fixed sample trial design? A case study by stochastic simulation. **J Clin Epidemiol** 1991; 44: 265–272.
15. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. **Biometrika** 1977; 64: 191–199.
16. Pasternack BS, Shore RE. Group sequential methods for cohort and case–control studies. **J Chron Dis** 1980; 33: 365–373.
17. Pasternack BS, Shore RE. Sample sizes for group sequential cohort and case–control study designs. **Am J Epidemiol** 1981; 113: 182–191.
18. Pasternack BS, Shore RE. Sample sizes for individually matched case–control studies: a group sequential approach. **Am J Epidemiol** 1982; 115: 778–784.

APPENDIX A

The logarithm of the likelihood ratio l_n is a function of δ, n ands u^2 and equal to

$$L = \ln(l_n) = \ln M(n/2;\ 1/2;\ \tfrac{1}{2}\cdot\delta^2\cdot u^2) - \tfrac{1}{2}\cdot n\cdot\delta^2. \quad \text{(A1)}$$

For the nth case–control pair ($n = 1, 2, 3, \ldots$ successively and one control observation per case) u^2 is equal to

$$u^2 = (\Sigma d_i)^2/\Sigma d_i^2 = n\cdot t^2/(n-1+t^2),\ i = 1, \ldots, n$$

where

$$t^2 = n\cdot \text{mean}(d)^2/\text{var}(d),$$

d_i is the difference between the observation for the case and the control observation, and mean(d) and var(d) stand for the mean and variance of these differences. For every n L is compared to $\ln(\beta/(1-\alpha))$ and $\ln((1-\beta)/\alpha)$. $M(a;b;x)$ is the confluent hypergeometric function, which can be calculated using Kummer's function [9], a series expansion:

$$M(a;b;x) = 1 + ax/b + a(a+1)x^2/\{b(b+1)2!\} + a(a+1)(a+2)x^3/\{b(b+1)(b+2)3!\} + \ldots$$

We involved 30 terms of this expansion. Rushton's approximation [8] to L is equal to

$$l_1 = \tfrac{1}{2}\cdot\delta\cdot u^3/\sqrt{n} + \sqrt{(n\cdot\delta^2\cdot u^2)} - (\tfrac{1}{2}\cdot n\cdot\delta^2 + \ln(2)). \quad \text{(A2)}$$

For k control observations per case the variance of the difference between the case observation and the mean of the k control observations is estimated using the cumulating case–control variance–covariance matrix. This estimate is then substituted as s^2 in the equations mentioned below. (The variance–covariance matrix takes the correlations among cases and controls into account. If we assume negligible correlations among control observations, equal variance for the control observations and equal correlations between the case and each of the controls, s^2 can be approximated by the variance of the differences between the case and the mean of the control observations.) Then Rushton's approximation to L can be calculated by

$$l_1 = \tfrac{1}{2}\cdot\delta\cdot u_k^3/\sqrt{n} + \sqrt{(n\cdot\delta^2\cdot u_k^2)} - (\tfrac{1}{2}\cdot n\cdot\delta^2 + \ln(2)) \quad \text{(A3)}$$

with

$$u_k^2 = n\cdot t_k^2/(n-1+t_k^2)$$

and

$$t_k^2 = n\cdot \text{mean}(d)^2/s^2.$$

N.B. For matched case–control observations ($k = 1$) equation (A3) is equal to equation (A2).

APPENDIX B

Data and calculations of one of the simulations with $\alpha = 0.05$, $1-\beta = 0.80$, $\delta = 0.5$, $\mu_0 = \mu_1 = 0.8$, $\sigma = 0.15$ and 2 controls per case are presented in Table B1 (see Appendix A for the notation used).

After 13 case–control sets are evaluated, M equals 1.002 and therefore $L = -1.623$ becomes smaller than the lower boundary, $\ln(\beta/(1-\alpha)) = -1.558$, and thus H_0 cannot be rejected.

When Rushton's approximation to L is applied, the sequential analysis can be stopped after the 10th case–control set, where $l_1 = -1.719$.

APPENDIX C

For matched case–control observations, the average sample number (ASN) for a sequential t-test with unknown variance is approxiamtely $(1+\delta^2/2)$ times the ASN for a test with known variance (Cox' approximation, Wetherill and Glazebrook [12]).

Under H_0 this ASN (unknown variance) is about

$$-2/\delta^2*\{\alpha'\cdot\ln((1-\beta)/\alpha') + (1-\alpha')\cdot\ln((\beta)/1-\alpha')\},$$

and under H_1 this ASN is about

$$(1+2/\delta^2)*\{\beta\cdot\ln((\beta)/(1-\alpha')) + (1-\beta)\cdot\ln((1-\beta)/(\alpha'))\}$$

(with $\alpha' = \alpha/2$). We recognize that Cox' approximation is an asymptotic result and that it is currently unknown how accurate it is.

Table B1

n	Case	Control	Control	s^2	t_2^2	u_2^2	M	L
1	0.911	0.912	0.891					
2	0.919	1.044	0.518	0.008	1.312	1.135	1.312	0.022
3	0.628	0.867	1.029	0.056	0.180	0.248	1.095	−0.284
4	0.781	0.759	0.861	0.037	0.273	0.334	1.174	−0.340
5	0.947	0.740	0.600	0.049	0.022	0.028	1.018	−0.608
6	0.527	0.728	0.791	0.050	0.084	0.099	1.075	−0.677
7	0.814	1.053	0.771	0.042	0.223	0.250	1.230	−0.668
8	0.784	0.730	0.877	0.036	0.263	0.290	1.308	−0.731
9	0.908	0.860	0.826	0.033	0.151	0.167	1.195	−0.947
10	0.745	0.637	0.580	0.032	0.018	0.020	1.025	−1.226
11	0.846	0.659	0.672	0.032	0.032	0.035	1.048	−1.328
12	0.650	0.762	0.919	0.032	0.019	0.020	1.031	−1.470
13	0.898	0.896	0.778	0.030	0.001	0.002	1.002	−1.623

Summary / Samenvatting / Résumé

Summary

Prospective cohort studies provide an ideal epidemiological approach to investigating the relation between diet, nutritional status, and cancer. For sufficient statistical power, however, which requires the observation of a minimum number of "cases" with disease, such prospective studies must usually comprise a very large number of individuals. As the costs of this type of study can be high, it is fundamental that an efficient design be used so that the study will be as informative as possible for a given investment of time and resources.

This thesis includes a series of methodological papers which present and analyse approaches that may improve the design and analysis of prospective cohort studies on diet, nutrition and chronic disease risk. The first chapters (2 to 5) examine methods of optimizing the assessment of the habitual, long-term dietary intake of cohort members, focusing on:

1. methods to maximize the amount of variation in true intake level of foods and nutrients that is actually distinguished – or "predicted" – by dietary questionnaire assessments collected at baseline; this is a means to increasing the power of a cohort study without increasing the number of study participants;
2. methods for the precise estimation of the distribution of predicted intake levels; this is essential for the accurate estimation of the statistical power or sample size requirements of the cohort, as well as for the unbiased estimation relative risks describing diet-disease associations.

Chapter 6, on the other hand, discusses the use of a sequential study design, to optimize the number of specific study hypotheses that can be evaluated when exposure assessments are based on a biochemical marker measured in urine, blood, or other tissue samples.

A first basic approach to maximize the predicted variation in true intake level is to select a dietary questionnaire method that allows an optimal classification of individuals by their respective intakes of foods and nutrients. Traditionally, the method is selected on the basis of the estimated correlation between questionnaire assessments and the individuals' true dietary intake levels. Chapter 2 reviews, in terms of latent variable models, how this correlation can be estimated in a preliminary validity

study, by comparison with at least two additional intake measurements. A vital assumption is that all measurements must have mutually independent errors, so that correlations between the measurements are entirely due to their relations with the same (latent) true intake variable. In practice, the additional measurements are most often obtained by means of repeated food intake records, using either a weighing method or 24-hour recalls. An alternative design of validity studies is presented, where the correlation between questionnaire assessments and true intake levels of nutrients is estimated by comparison with food intake records as well as with a biochemical marker. The advantage of this alternative approach is that measurement errors are more likely to be independent when all three measurements are taken with different methods (i.e., questionnaire, records, and biochemical marker).

The estimated coefficient of correlation between questionnaire assessments and the individuals' true habitual intakes is also seen as essential information for the subsequent planning of epidemiological studies on diet:

a) to evaluate the sample size requirements of a cohort study with correction for power losses due to random errors in the dietary exposure assessments; and

b) to estimate the magnitude of attenuation bias in relative risks.

In **Chapter 3** it is shown that, if relative risks are estimated for scaled, absolute differences in intake level (expressed in standard units), sample size requirements for a cohort study can be computed from the variance of the distribution of true dietary intake levels predicted by the questionnaire assessments collected at baseline. Likewise, bias in relative risk estimates can be shown to be equal to the variance of the predicted intake distribution divided by the variance of the questionnaire assessments. To estimate the variance of the predicted intake distribution, it is not necessary to know the correlation between questionnaire assessments and true dietary intake values. Thus, a validity study based on multiple additional measurements is not essential. Instead, a calibration study can be used, based on only a single day's food intake record per person as a reference measurement. It is shown that, for a given total number of daily intake records taken, the estimation of the variation in predicted intake levels is most precise when the calibration is based on a

maximum number of participants, with only a single record each. An additional major advantage is that a calibration study can be conducted more easily on a representative sample of the study population (when nested within a prospective cohort study). Representativeness is an important condition for accurate estimation of the power of cohort studies, or of biases in relative risk estimates.

A second approach to increase the magnitude of the predicted variation in intake values is to broaden the range of true dietary intake levels covered. This can be achieved by combining the data from multiple cohort studies, conducted in different geographical areas with heterogeneous life styles and dietary habits. An example of this multi-cohort approach is the EPIC study (European Prospective Investigation on Cancer and Nutrition), which is coordinated by the International Agency for Research on Cancer at Lyon (France). In **Chapter 4** it is shown that in such multi-cohort projects, relative risks indicating associations between dietary intake level and disease incidence can be estimated from:

1. within-cohort differences in the measured dietary intake levels and disease outcomes of individuals; and
2. between-cohort ("ecological") variation, between mean intake measurements and mean incidence rates of disease at a population level.

If there is sufficient concordance between the various component estimates, these can be combined into an overall, more powerful summary value. A complication, however, is that relative risk estimates within different cohorts can be biased to various degrees as a result of dietary assessment errors, while the between-cohort ("ecological") relation may also be distorted by differences in systematic over- or under-estimation of mean intake levels. This may be particularly true when it is impossible to use an identical method for dietary intake assessment in all cohorts. The second part of Chapter 4 proposes the use of sub-studies for the calibration of dietary intake assessments to adjust for possible heterogeneity in relative risk estimates due to such divergent biases. This reduction in heterogeneity may improve the power of a statistical test for diet-disease association based on a pooled estimate of relative risk.

Calibration adjustments for biases in relative risk estimates will only be adequate if the calibration factors used for such corrections are themselves estimated with sufficient precision. This aspect is discussed in

Chapter 5, which presents a simplified approach to the estimation of approximate sample size requirements for dietary calibration studies nested within a cohort. These sample size estimations are based on two alternative criteria, requiring either a minimum relative efficiency of calibration (so that there is little loss of precision in the estimation of relative risk), or a minimum statistical power of a test for diet-disease association based on the corrected relative risk estimate (i.e., after calibration). The required size of a calibration study then depends only on the correlation between questionnaire assessments and reference measurements.

In studies where the exposure assessments are based on a biochemical marker, measured in blood, urine, or other biological specimens, a simple efficiency measure is to store the biological specimens in a "biobank", and to postpone the exposure measurement until it is known which individuals develop a given type of disease, and which will be suitable control subjects. Nevertheless, the number of scientific hypotheses potentially of interest is usually much larger than the number of biomarkers that can be actually assessed with a limited amount of blood or other biological specimens available. It would thus be useful to have a statistical method which, at the expense of as little biological material as possible, distinguishes between promising or less promising hypotheses. For this purpose, Chapter 6 proposes the use of a sequential study design, in which laboratory analyses of the biological specimens of cases and controls are conducted until sufficient data have accumulated to either reject or not a null hypothesis of "no association" between marker and disease risk. On average, as compared to an equivalent fixed-sample test procedure, a sequential test may require less than half the number of biological specimens to reach a conclusion. If the null hypothesis is rejected, additional biological specimens may be analyzed to improve the precision of relative risk estimates; if not, biological specimens can be spared for the evaluation of different hypotheses.

In conclusion, preliminary validity studies in which the correlation between questionnaire assessments and true dietary intake levels is estimated, may be used to select an optimal questionnaire instrument to be employed in a prospective cohort study. On the other hand, calibration studies using only one reference measurement per person are more efficient

when the objective is to estimate the power of prospective cohort studies, accounting for the effects of random dietary assessment errors, or to correct for biases in relative risk estimates. A main advantage of calibration studies is that these can be conducted more easily on a representative sample of the study population. In multi-cohort projects, calibration studies can be used to improve the comparability of cohort-specific relative risk estimates, and to obtain a more precise estimate of the between-cohort, "ecological" relation between dietary intake levels and disease incidence. In studies where the exposure assessments are based on a biochemical marker, a simple efficiency measure is to store biological specimens in a biobank, and to postpone laboratory analyses until cases with disease have been identified. Sequential study designs can then be used to allow the evaluation of an maximum number of scientific hypotheses with a given amount of biological material available.

Samenvatting

Prospectieve cohort studies bieden een ideale epidemiologische benadering om relaties tussen voeding, voedingstoestand and kanker te bestuderen. Echter, om een voldoende groot statistisch onderscheidingsvermogen te ontwikkelen – hetgeen vereist dat een minimum aantal ziektegevallen wordt waargenomen – moeten prospectieve studies over het algemeen een groot aantal individuen omvatten. De kosten van dit type studie kunnen daarom hoog oplopen, en het is dus van fundamenteel belang de studie efficient op te zetten, zodat zoveel mogelijk informatie wordt verkregen voor een gegeven investering in tijd en middelen.

Dit proefschrift bevat een reeks methodologische artikelen waarin een aantal benaderingen worden gepresenteerd en besproken om de opzet en analyse van prospectieve cohort studies over voeding, voedingstoestand en chronische ziekten te verbeteren. In de eerste hoofdstukken (2 tot 5) worden methoden onderzocht om een optimale meting te verkrijgen van de gebruikelijke voedings inname op langere termijn van individuen in de cohort study. Hierbij wordt de nadruk gelegd op:

1. methoden om de variatie in innameniveau van voedingsmiddelen of nutrienten die werkelijk wordt onderscheiden – ofwel "voorspeld" – door metingen aan het begin van de studie zo groot mogelijk te maken; dit is een manier om het onderscheidingsvermogen van een cohort study te doen toenemen zonder het aantal deelnemers in de studie te vergroten;
2. methoden voor een nauwkeurige schatting van de werkelijk gemeten variatie in innameniveaus; dit is essentieel voor een nauwkeurige schatting van het onderscheidingsvermogen of de vereiste steekproefgrootte van het cohort, zoel als voor de zuivere schatting van relatief risiko's die verbanden tussen voeding en ziekte beschrijven.

Hoofdstuk 6 bespreekt de toepassing van een sequentiële studie-opzet om een optimaal aantal specifieke hypothesen te kunnen toetsen wanneer potentiele risikofactoren worden gemeten in urine, bloed, of biologische weefselmonsters.

Een eerste basisbenadering om de werkelijk gemeten ("voorspelde") variatie in innameniveau zo groot mogelijk te maken is het selecteren van

een vragenlijst methode die leidt tot een tot een optimale classificatie van individuen naar hun gebruikelijke inname van voedingsmiddelen of nutrienten. Deze selectie wordt traditioneel gemaakt op basis van de geschatte correlatie tussen inname-metingen verkregen via de vragenlijst en werkelijke inname-niveaus. **Hoofdstuk 2** geeft een overzicht, in termen van "latent variable" modellen, van benaderingen om deze correlatie te schatten in een voorafgaande validatie-studie. Dit vereist een door vergelijking met tenminste twee extra innamemetingen. Daarbij is het een essentiele aanname dat fouten in de verschillende metingen wederzijds onafhankelijk zijn, zodat correlaties tussen de metingen uitsluitend het gevolg zijn van hun relaties met dezelfde (latente) inname variabele. In de praktijk worden de extra metingen meestal verkregen met behulp van een meerdaagse gewogen opschrijfmethode, of door middel van herhaalde 24-uurs recalls. In hoodstuk 2 wordt ook een alternatieve opzet van validatie-studies gepresenteerd waarin de correlatie tussen vragenlijstmetingen en werkelijke innameniveaus van nutrienten wordt geschat door vergelijking met metingen verkregen via een opschrijfmethode zowel als met metingen gebaseerd op een biochemische parameter. Het voordeel van deze laatste benadering is dat de aanname van onafkankelijke meetfouten gemakkelijker kan worden gemaakt als alle drie metingen worden verkregen via verschillende methoden (d.w.z., vragenlijst, opschrijfmethode of 24-uurs recall, en biochemische parameter).

De geschatte correlatie tussen vragenlijst-metingen en werkelijke innameniveaus wordt ook gezien als essentiële informatie voor de verdere planning van epidemiologische studies met betrekking tot de voeding:

a) *om de vereiste steekproefgrootte van een cohort studie te schatten*, daarbij rekening houdend met het verlies aan onderscheidingsvermogen als een gevolg van toevallige (d.w.z., "random") fouten in de voedingsinname-metingen; en
b) om de grootte van attenuatie bias in relatieve risiko's te schatten

In **hoofdstuk 3** wordt getoond dat, als relatieve risiko's worden geschat voor absolute verschillen in inname-niveau, uitgedrukt in standaard eenheden, de vereiste steekproefgrootte voor een cohort studie kan worden berekend uit de variantie van de verdeling van werkelijke innamewaarden zoals die worden voorspeld door vragenlijst-metingen verkregen aan het begin van de studie. Bovendien blijkt de de bias in relatief risiko-schattingen gelijk te zijn aan de variantie van de verdeling van voorspelde innamewaarden gedeeld door

de variantie van vragenlijstmetingen. Om de variantie van de voorspelde innamewaarden te schatten is het niet noodzakelijk de correlatie tussen vragenlijstmetingen en werkelijke innamewaarden te kennen. Een validatie-studie gebaseerd op meer dan één aanvullende innamemeting is dus niet echt vereist. In plaats daarvan kan een calibratiestudie worden opgezet, gebaseerd op een slechts ééndaagse gewogen opschrijfmethode, of een enkele 24-uurs recall. Het wordt getoond dat, voor een gegeven totaal aantal dagelijkse innamemetingen, de schatting van de voorspelde variatie in innameniveaus het meest nauwkeurig is wanneer de calibratiestudie een maximum aantal deelnemers omvat, met ieder slechts één enkele innamemeting. Een groot voordeel van deze benadering is dat een caliebratie-studie gemakkelijker kan worden uitgevoerd in een representatieve steekproef van de onderzoekspopulatie (indien genest in een prospectief cohort onderzoek). Representativiteit is een belangrijke voorwaarde voor een correcte schatting van het onderscheidingsvermogen van een cohort-studie, of van bias in schattingen van relatieve risiko's.

Een tweede benadering om de voorspelde variatie in de innamewaarden van individuen te vergroten is het bereik van werkelijke innameniveaus te verbreden. Dit kan worden bereikt door het combineren van gegevens verkregen in meerdere cohort-onderzoeken, uitgevoerd in geografische gebieden met verschillende leefstijlen en voedingsgewoonten. Een voorbeeld van deze multi-cohort benadering is de EPIC-studie (European Prospective Investigation on Cancer), die wordt gecoordineerd door de International Agency for Research on Cancer in Lyon (Frankrijk). Hoodstuk 4 laat zien dat in dit type onderzoek relatieve risiko's die het verband aangeven tussen voedingsinname-niveaus en de incidentie van ziekte kunnen worden geschat uit:

1. binnen-cohort variatie in gemeten innameniveaus en ziekte-uitkomst van afzonderlijke individuen; en
2. tussen-cohort variatie in de gemiddelde innamemetingen en ziekte-incidenties op populatie-niveau.

Als er voldoende overeenkomst is tussen deze verschillende relatief risiko-schattingen dan kunnen deze worden verenigd in een samenvattende waarde met een grotere precisie. Een complicatie hierbij is echter dat fouten in de voedingsinname-metingen in ongelijke mate bias kunnen geven aan relatief rissiko-schattingen verkregen in verchillende cohorten, terwijl ook de

tussen-cohort ("ecologische") relatie tussen voedingsinname en ziekterisiko kan worden verstoord door systematische over- en onder-schattingen van innameniveaus. Dit kan met name het geval zijn wanneer het niet mogelijk eenzelfde methode te gebruiken voor het meten van de voedingsinname in alle cohorten. In het tweede deel van **hoofdstuk 4** wordt voorgesteld om sub-studies voor de calibratie van voedingsinname-metingen te gebruiken voor de correctie van variatie in relatief risiko-schattingen als gevolg van verschillen in bias. De reductie in de bias in de verschillende relatief risiko-schattingen kan zo het onderscheidingsvermogen van een statistische toets voor een verband tussen voeding en ziekte vergroten, wanneer deze toets is gebaseerd op een samenvattende relatief risiko-schatting in een multi-cohort onderzoek.

Calibratie-correcties voor bias in relatief risiko-schattingen zullen alleen doeltreffend zijn als de gebruikte calibratie-factoren zelf met voldoende precisie worden geschat. Dit aspect wordt besproken in **hoofdstuk 5**, waarin een eenvoudige methode wordt gepresenteerd voor het schatten van de vereiste steekproefgrootte van calibratie-studies binnen een cohort. Deze steekproefgrootte-schattingen zijn gebaseerd op twee alternatieve criteria. Deze criteria vereisen een minimale relatieve efficiëntie van calibratie (d.w.z., zodat er slechts weinig verlies in precisie in relatief risikoschattingen optreedt), ofwel een minimum onderscheidingsvermogen van een statistische toets voor een verband tussen voeding en ziekte, als deze toets is gebaseerd op de gecorrigeerde (d.w.z. "gecalibreerde") relatief risiko-schatting. De benodigde steekproefgrootte van een calibratiestudie hangt dan uitsluitend af van de correlatie tussen vragenlijst- en referentie-metingen.

In studies waar een potentiële risiko-factor wordt gemeten in urine, bloed, of andere biologische monsters (d.w.z., in de vorm van een "biomarker"), kan de efficiëntie van de studie op eenvoudige wijze worden verbeterd door de biologische monsters op te slaan in een "biobank", en door meting van de risikofactor uit te stellen tot het bekend is welke individuen een bepaald type ziekte hebben ontwikkeld, en wie daarbij als controle-personen kunnen worden geselecteerd. Desondanks is het aantal te toetsen hypothesen meestal veel groter dan het aantal biomarkers dat kan worden gemeten in de beschikbare hoeveelheid bloed, of andere biologische monsters. Het is daarom van belang over een statistische methode te beschikken om, met

gebruik van zo weinig mogelijk biologisch materiaal, onderscheid te maken tussen veelbelovende en minder interessante hypothesen. Voor dit doel wordt in **hoofdstuk 6** een sequentiële onderzoeksopzet voorgesteld, waarin laboratoriumanalyses van biologische monsters van cases met ziekte en controlepersonen net zolang worden uitgevoerd tot er voldoende gegevens zijn om de nulhypothese van "geen verband" tussen biomarker en ziekterisiko al of niet te verwerpen. Vergeleken met een statistische toets gebaseerd op een vaste steekproefomvang, kan via deze sequentiële benadering minder dan de helft van het aantal laboratoriumanalyses volstaan om tot een conclusie te komen. Als de nulhypothese wordt verworpen kunnen extra biologische monsters worden geanalyseerd om de precisie van relatief risiko-schattingen te verbeteren; zo niet, dan kunnen biologische monsters worden gespaard voor het toetsen van andere hypothesen.

Tot besluit, validatiestudies waarin de correlatie tussen vragenlijst-metingen en werkelijke voedingsinname-niveaus worden geschat kunnen worden gebruikt om een optimale vragenlijstmethode te selecteren voor toepassing in een prospectief cohortonderzoek. Aan de andere kant zijn calibratiestudies met slechts één referentiemeting per persoon efficiënter voor het schatten van het statistisch onderscheidingsvermogen van een cohortonderzoek – daarbij rekening houdend met de effecten van toevalsfouten in innamemetingen – of voor de correctie van bias in relatief risiko-schattingen. Een belangrijk voordeel van calibratiestudies is dat deze gemakkelijker kunnen worden uitgevoerd in a representatieve steekproef van de onderzoekspopulatie. In multi-cohort projecten kunnen calibratiestudies worden gebruikt om de vergelijkbaarheid van relatief risiko-schattingen te verbeteren, en om een nauwkeurigere schatting the verkrijgen van de "ecologische" relatie tussen gemiddelde voedingsinnamen en ziekterisiko's in verschillende cohorten. In studies waar een risikofactor wordt gemeten met behulp van een biomarker, kan de efficientie eenvoudig worden vergroot door biologische monsters op te slaan in een biobank, en laboratorium-analyses uit te stellen totdat "cases" met ziekte zijn geïdentificeerd. Gebruik van een sequentiële studie-opzet maakt het dan mogelijk een optimaal aantal wetenschappelijke hypothesen te evalueren.

Résumé

Les études prospectives de cohorte constituent une approche idéale pour étudier les relations entre le régime, le statut nutritionnel et le cancer. Toutefois, pour obtenir une puissance statistique suffisante, ce qui exige l'observation d'un minimum de "cas" atteints de la maladie, de telles études prospectives doivent en général englober un très grand nombre d'individus. Le coût d'une telle étude peut donc être élevé. C'est pourquoi il est capital de choisir un protocole efficace qui rende l'étude aussi informative que possible pour un investissement temps/ argent donné.

La présente thèse contient une série de conseils méthodologiques qui proposent et analysent des démarches qui pourraient améliorer la conception et l'analyse des études prospectives de cohorte sur l'alimentation, la nutrition et le risque de maladie chronique. Les premiers chapitres (2 à 5) passent en revue des méthodes qui pourraient optimiser l'évaluation de la consommation alimentaire habituelle à long-terme des membres de la cohorte en attirant l'attention sur :

1. les moyens de maximaliser la quantité de variations dans la consommation réelle d'aliments et de nutriments qui est en fait évaluée -ou "prédite"- par le biais des bilans alimentaires recueillis à la base. Ceci est une des méthodes permettant d'augmenter la puissance d'une étude de cohorte sans pour autant augmenter le nombre des participants à cette étude ;
2. les méthodes permettant d'estimer la distribution des niveaux de consommation prédits de façon précise ; ceci est indispensable pour obtenir une évaluation exacte de la puissance statistique ou de la taille nécessaire de l'échantillon de la cohorte ainsi que pour une estimation objective des risques relatifs décrivant les associations entre la maladie et les habitudes alimentaires.

Quant au Chapitre 6, il traite de l'utilisation d'un protocole d'étude séquentielle dans le but d'optimiser le nombre d'hypothèses d'études spécifiques qui peuvent être obtenues lorsque l'on évalue l'exposition à partir de marqueurs biochimiques mesurés dans les urines, le sang ou d'autres échantillons de tissu.

Une première approche simple permettant de maximaliser la variation prédite en niveau réel de consommation consiste en la selection d'un type

de questionnaire alimentaire qui autorise la classification optimale des individus en fonction de leur consommation respective d'aliments et de nutriments. Traditionnellement, on fait cette sélection en se basant sur les corrélations estimées entre les bilans alimentaires et le niveau de consommation réel des individus. Le **Chapitre 2** examine, en termes de modèles de variable latente, comment estimer cette corrélation au cours d'une étude préliminaire de validité, en comparaison avec au moins deux mesures de consommation. Il est primordial que toutes les mesures comportent des erreurs indépendantes les unes des autres afin que la corrélation entre les mesures soient uniquement due à leur relation avec la même variable de consommation réelle (latente). En pratique, les mesures complémentaires sont le plus souvent obtenues au moyen d'enregistrements de consommation répétés qui utilisent soit une méthode de pesée, soit un rappel de 24 heures. On peut aussi concevoir des études de validité où la corrélation entre le questionnaire d'évaluation et les niveaux de consommation réels de nutriments seraient estimés en les comparant aux enregistrements de consommation alimentaire et aux marqueurs biochimiques. Cette alternative présente l'avantage que les erreurs de mesures ont plus de chances d'être indépendantes quand les trois mesures sont obtenues par des méthodes différentes (c'est à dire, questionnaire, enregistrement, et marqueurs biochimiques).

Les coefficients de corrélation estimés entre les évaluations obtenues par questionnaire et la consommation habituelle réelle de l'individu constituent aussi une information essentielle pour la plannification ultérieure d'études épidémiologiques sur l'alimentation :

a) pour évaluer la taille nécessaire de l'échantillon d'une étude de cohorte avec les corrections pour les pertes de puissance dues aux erreurs alléatoires dans l'évaluation de l'exposition alimentaire ; et
b) pour calculer une estimation de l'amplitude du biais d'atténuation dans le risque relatif.

Dans le **Chapitre 3**, on démontre que si les risques relatifs sont calculés pour des différences échelonnées et absolues dans le niveau de consommation alimentaire (exprimées en unités standard), on peut calculer la taille requise de l'échantillon à partir de la variance de distribution des niveaux de consommation réels prédits par les questionnaires d'évaluation recueillis

au départ. De même, on peut démontrer que le biais des risques relatifs estimés est égal à la variance de la distribution de la consommation prévue divisée par la variance du questionnaire d'évaluation. Il n'est pas nécessaire de connaître la corrélation entre les questionnaires d'évaluation et les valeurs de consommation réelle pour évaluer la variance de la distribution de la consommation prédite. C'est pourquoi une étude de validité basée sur de multiples mesures complémentaires n'est pas indispensable. On peut utiliser à la place une étude de calibrage qui utilise comme mesure de référence la consommation en aliments d'un seul jour pour une seule personne. On sait que pour un nombre donné d'enregistrements sur la consommation journalière, l'estimation de la variation des niveaux de consommation prédits est la plus précise quand le calibrage est basé sur un nombre maximum de participants, avec seulement un enregistrement chacun. Autre avantage important : l'étude de calibrage peut être mise en place plus facilement sur un échantillon représentatif de la population de l'étude (quand il est inclus dans une étude prospective de cohorte). La représentativité est une condition importante pour une estimation exacte de la puissance des études de cohorte, ou des biais d'estimation du risque relatif.

La deuxième démarche permettant d'augmenter l'amplitude des variations prédites des valeurs de consommation est d'élargir la gamme de consommation alimentaire réelle couverte. On peut obtenir cela en combinant les données de différentes études de cohorte, réalisées dans des régions du monde différentes et recouvrant des styles de vie et des habitudes alimentaires différents. Le programme EPIC (étude prospective de recherche sur le cancer en Europe, coordonnée par le Centre International de Recherche sur le Cancer (Lyon, France)) est un exemple de cette approche multi-cohortes. Le **Chapitre** 4 montre que dans les projets multi-cohortes de ce genre, le risque relatif indiquant des associations entre le niveau de consommation alimentaire et l'incidence de la maladie peut être estimé à partir :

1. des différences internes aux cohortes dans les niveaux de consommation alimentaires mesurés et les pronostics de la maladie chez les individus ;
2. des variations ("écologiques") inter-cohortes, entre les mesures de consommation moyenne et les taux moyens d'incidence de la maladie au niveau de la population.

Si la cohérence entre les diverses estimations des composés est suffisante,

on peut les rassembler dans un résumé des valeurs générales plus puissant. Il est toutefois préférable de rester prudent puisque, à l'intérieur des différentes cohortes, les estimations des risques relatifs peuvent être biaisées à différents degrés à la suite d'erreurs dans les bilans alimentaires, de même que les relations inter-cohortes ("écologiques") peuvent aussi être faussées par des différences entre les inévitables "sur" ou "sous-évaluations" des niveaux de consommation moyens. Ceci est particulièrement vrai lorsqu'il est impossible d'utiliser des méthodes identiques pour évaluer la consommation alimentaire dans toutes les cohortes. La deuxième partie du Chapitre 4 propose d'utiliser des mini-études pour le calibrage des bilans alimentaires permettant un ajustement pour l'hétérogénéité possible dans les risques relatifs estimés, dont les biais de divergeance sont responsables. Cette diminution de l'hétérogénéité dans les risques relatifs estimés pourrait augmenter la puissance d'un test statistique basé sur les risques relatifs estimés groupés et permettant d'établir la relation entre alimentation et maladie.

Les ajustements de calibrage pour les biais de risques relatifs ne seront appropriés que si les facteurs de calibrage utilisés pour ces corrections ont eux-mêmes été établis avec suffisamment de précision. Cet aspect du problème est abordé dans le **Chapitre 5** qui présente une approche simplifiée de l'estimation de la taille approximative nécessaire à un échantillon au sein d'une cohorte. Ces estimations de la taille d'un échantillon sont basées sur deux critères alternatifs, qui exigent soit une capacité de rendement relative minimum du calibrage (pour qu'il n'y ait qu'une perte minime de précision dans l'estimation du risque relatif), soit un test doté d'un minimum de puissance statistique pour établir une relation entre l'alimentation et la maladie basée sur les estimations corrigées des risques relatifs (c'est à dire, après calibrage). La taille nécessaire d'une étude de calibrage ne dépend alors plus que de la corrélation entre les évaluations obtenues par questionnaires et les mesures de référence.

Dans les études où les évaluations de l'exposition sont basées sur des marqueurs biochimiques mesurés dans le sang, les urines ou d'autres spécimens biologiques, une mesure performante et simple est de stocker les échantillons biologiques dans une "biobanque", et de repousser la mesure de l'exposition jusqu'à ce qu'on sache quels sont les individus qui ont eu un

type donné de maladie et qui fera un contrôle approprié. Il n'en reste pas moins que le nombre d'hypothèses scientifiques qui présentent un intérêt potentiel est généralement bien plus grand que le nombre de marqueurs biologiques que l'on peut effectivement mesurer avec une quantité limitée de sang ou d'un autre échantillon biologique disponible. Il serait donc utile de disposer d'une méthode statistique qui, avec le moins de matériau biologique possible, permette de distinguer les hypothèses les plus interessantes des moins intéressantes. C'est dans ce but que le **Chapitre 6** propose d'utiliser un plan d'étude séquentiel dans lequel les analyses des laboratoires des échantillons biologiques provenant des cas et des contrôles seraient effectuées jusqu'à ce que l'on dispose de suffisamment de données pour rejeter ou conserver une hypothèse nulle de "non association" entre marqueur et risque de maladie. En moyenne, et en comparaison à une procédure équivalente de test sur échantillon fixe, un test séquentiel demande moins de la moitié des échantillons biologiques pour aboutir à une conclusion. Si l'hypothèse nulle est rejetée, on peut analyser des échantillons biologiques suplémentaires pour améliorer la précision des risques relatifs estimés ; sinon, on peut garder les échantillons biologiques pour évaluer les différentes hypothèses.

On peut donc conclure que les études préliminaires de validité qui évaluent la corrélation entre les bilans et la consommation alimentaire réelle, peuvent servir à selectionner le type de questionnaire optimal qui pourra être utilisé dans une étude prospective de cohorte. D'autre part, les études de calibrage qui n'utilisent qu'une mesure de référence par personne sont plus efficaces quand leur objectif est d'estimer la puissance des études prospectives de cohortes en rendant compte des effets causés par les erreurs d'évaluation aléatoires ou de corriger des biais dans les risques relatifs estimés. Le principal avantage des études de calibrage est qu'elles peuvent être effectuées plus facilement sur un échantillon représentatif de la population de l'étude. Dans le cadre des projets multi-cohortes, les études de calibrage peuvent être utilisées pour améliorer la qualité de comparaison entre les risques relatifs estimés de chaque cohorte et pour obtenir une estimation plus précise de la relation inter-cohorte, "écologique" entre les niveaux de consommation alimentaires et l'incidence de la maladie. Dans les études où les évaluations de l'exposition sont basées sur les marqueurs biochimiques, une mesure simple et efficace est de

stocker les échantillons biologiques dans une "biobanque" et de repousser les analyses de laboratoire jusqu'à ce qu'on ait rencontré des cas de maladie. On peut ensuite utiliser les plans d'études séquentielles pour permettre l'évaluation d'un nombre optimal d'hypothèses scientifiques avec une quantité donnée de matériau biologique disponible.

Acknowledgements

This thesis has been written with the support of many colleagues and friends. First of all, I would like to mention the various national investigators participating in the EPIC steering commitee meetings. It was during these meetings that we developed the idea of a dietary "calibration" approach, which has become a central theme of this thesis. I am particularly indebted to Martyn Plummer and David Clayton who, at the several times they invited me to stay with them at Cambridge, have generously shared their insights in statistical aspects of validation and calibration of dietary intake measurements. Without them this thesis would have never had its present form. Likewise, I have appreciated highly the collaboration with Ingeborg van der Tweel on the preparation of Chapter 6 on the use of sequential methods.

The support of my two official supervisors – Wija van Staveren and Elio Riboli – has of course been very important: they kept me on the right track towards the completion of this thesis. I also feel most obliged to Elio for the way he kept his team at work whenever the external funding of the EPIC project was late due to bureaucratic obstacles. Needless to say that he created the necessary conditions for the realization of this work.

Among my colleagues at the IARC, I want to thank first of all Sarah Somerville for her editorial assistance. I am also grateful to Jacques Estève, Annie Sasco, Paul Demers, and Christine Friedenreich, for critical comments on draft papers, and to Corinne Casagrande, Bertrand Hémon, Anne Linda van Kappel, and Cathérine Gros for their help with computer analyses, simulations, and graphics (which, if not always retained for the final text of this thesis, have been of great value to improve our understanding of calibration problems). I shall always remember Nadia Slimani for the philosophical discussions we had during work and at lunch.

Finally, I want to thank Franca for her being so patient (a difficult exercise !), during moments of my physical or mental absence. Now that the little "formality" of this thesis is over, we'll hopefully find more time to think about other important things in life. And of course I am particularly thankful to my parents, to whom a good education of their children has always been of principle concern. They should consider this booklet as dedicated to them.

Affiliations of co-authors

Jacques Estève Ph.D., statistician.
International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cédex 08, France (Tel: 72 73 84 85. Fax: 72 73 85 75).

Rudolf Kaaks M.Sc., nutritional epidemiologist.
Programme of Nutrition and Cancer, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cédex 08, France.

Martyn Plummer M.Sc., statistician.
MRC Biostatistics Unit, Institute of Public Health, University Forvie Site Robinson Way Cambridge CB2 2SR, United Kingdom.

Elio Riboli M.D., M.Sc., epidemiologist.
Head, Programme of Nutrition and Cancer, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cédex 08, France.

Anne Linda van Kappel M.Sc., nutritionist.
Programme of Nutrition and Cancer, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cédex 08, France.

Paul van Noord M.D., Ph.D., epidemiologist.
Department of Public Health and Epidemiology, University of Utrecht, Radboudkwartier 261-263, 3511 CK Utrecht, the Netherlands.

Ingeborg van der Tweel, statistician.
Centre for Biostatistics, University of Utrecht, Centrumgebouw Noord C122, Padualaan 14, 3584 CH Utrecht, the Netherlands.

Wija van Staveren Ph.D., professor geriatric nutrition.
Department of Human Nutrition, Wageningen Agricultural University, P.O. Box 8129, 6703 BC Wageningen, The Netherlands.

Curriculum vitae

Rudolf Johan Kaaks was born on the 27th of March 1960, in Groningen, the Netherlands. In 1978 he finished secondary school at the "Stedelijk Gymnasium" (β-orientation) at Nijmegen. In September 1978 he started his studies at the Agricultural University at Wageningen, where he graduated in Nutrition Sciences in January 1987 (majors: nutrition, epidemiology, and statistics). During his studies he received a special training award from the International Agency for Research on Cancer, Lyon, to work during nine months on the analysis of two case-control studies on diet and gastro-intestinal cancers. From 1987 to October 1988 he worked at the University of Utrecht (Department of Epidemiology), investigating relations between risk factors for breast cancer and the Dutch famine during the Second World War. In October 1988, he returned to the International Agency for Research on Cancer, to continue work on case-control studies, and to assist in the development of a European, multi-centre cohort study on diet, nutrition, and cancer (now well established as the "EPIC" project). While working on this project, he started the preparation of his PhD thesis in early 1990. In the meantime he has developed a strong interest in biological mechanisms relating diet, metabolic (hormonal) status, and cancer, towards which he hopes to direct his future work at the IARC.