EDUCATING STATISTICIANS FOR WORK IN EPIDEMIOLOGY: CAN WE FIND THE CORRECT BALANCE BETWEEN GENERAL STATISTICAL ABILITY AND PARTICULAR SKILLS?

Ivar Heuch
University of Bergen
Norway

The comprehensive population registries available in Scandinavia have made possible extensive work in epidemiology on associations between risk factors and disease. The area has attracted many statisticians with little epidemiological training. Experience has shown that some find it difficult to adapt to the practical challenges. Not only is a basic understanding required of statistical methods, but a particular cautious attitude is needed in interpretation of epidemiological data. The statistician must often deal with issues of a biological nature, as well as data processing problems. All aspects must be taken into account in the education of professional statisticians. It is argued that the non-statistical components should be integrated into the training, with a considerable freedom of choice for the students. If relevant data analytic work is included, students will be exposed to many of the challenges encountered in epidemiology.

INTRODUCTION

The field of epidemiology deals with the distribution of disease in populations, as well as underlying risk factors and their relation to disease, studied on the basis of population data. Biostatistical methods play an important role in epidemiology, although the observational nature of the data differs from standards in other disciplines of biostatistics such as clinical trials. Work in epidemiology has traditionally been carried out by professionals trained in medicine and public health, although many epidemiologists are nowadays recruited from demography, anthropology, behavioural sciences and statistics (Buffington et al., 1999). The interaction between persons with different skills has probably contributed to the development of this area. Historically, many important methodological advances have been made by statisticians working in epidemiology (Ellenberg et al., 1997). The use of complex statistical methods has increased dramatically in epidemiology over recent decades (Levy and Stolte, 2000).

The Scandinavian countries offer special opportunities for carrying out large, population-based studies in epidemiology. By means of personal identification numbers, individuals can be traced through all stages of life. Information from various sources on potential risk factors and disease can in principle be linked, although problems related to confidentiality sometimes make this process more difficult. Particular registries are available which cover the entire population with regard to special diseases, e.g. cancer. Thus it is relatively easy to organise studies in Scandinavia which include information collected over a long period.

Graduates in statistics from the science faculties at the universities in Norway have found jobs in a wide range of applied areas. Depending on the general economic situation, research institutions and industry have often offered the best opportunities. In recent years, insurance has been the most popular area of employment for statistics graduates from University of Bergen, followed by medical research. Most of the jobs associated with medicine have been in epidemiology. The experiences made by the graduates have been quite different, however. In some cases, they have clearly achieved great success working within this area. Other graduates have preferred to move into completely different areas of employment after some time.

In this paper I shall discuss what kind of skills are needed among statisticians working in epidemiology, as seen from the point of view of someone associated with the general training of statisticians. To the extent that my conclusions are based on actual observations, it must be taken for granted that the data collection may be severely biased. Nevertheless, it seems useful to formulate some principles which should be kept in mind in future reforms of the university education. Some of the recommendations will also be relevant outside the area of epidemiology.

THE NORWEGIAN BACKGROUND

The education of statisticians at Norwegian universities, at faculties of mathematics and natural sciences, is organised with a first degree, corresponding roughly to a B.Sc., after 3.5 years, and a second degree, similar to an M.Sc., after another 1.5 years. The lower degree includes mathematics, computer science, statistics, and other subjects chosen by the student, frequently economics. The higher degree usually comprises statistics only. Few students finish with the lower degree. A substantial part of the higher degree consists of writing a thesis, parts of which can be practical, although a major proportion is expected to be methodological. A general reform of the university system is on its way after political decisions made at the top level, but little is known about details.

Although students may include courses in almost any subject in the first degree, it is in practice difficult to find sufficient space. Basic biology is one possibility. Medicine and more specifically epidemiology belong to a different university faculty and courses in such areas are not easily included. Many students who aim at a degree in statistics, have no definite idea about future work, although some express general preferences.

My Norwegian experiences form the basis of the following discussion, but the general ideas should also apply to university degrees organised in a different way.

STATISTICAL SKILLS REQUIRED

Work in epidemiology requires a similar basic training of statisticians as jobs in many other applied areas, but there are some special requirements. Statistical problems present themselves in epidemiology at several levels of complexity. Fundamental questions concerning, for example, selection bias in samples, often lead to problems in probability calculus which are quite simple from a mathematical point of view. The student must still realise that not every problem involving selection bias can be solved in the same simple way.

A great many general concepts introduced at a secondary level in statistics are constantly referred to in epidemiology. One does not get very far without familiarity with topics such as maximum likelihood estimation and likelihood ratio tests. More theoretical concepts as sufficient statistics occur frequently. The statistician is often faced with a choice between epidemiological methods supported by different general principles. In many situations, especially with stratification, problems arise in models with potentially irrelevant nuisance parameters, which may be handled through various kinds of conditioning. The essential point is that the statistician working in epidemiology must be familiar with such fundamental concepts in connection with statistical inference, but not necessarily with the complete theoretical foundation.

Particular methods which find widespread use in epidemiology must also be considered. Logistic regression and generalised linear models, survival analysis and more generally life history analysis are important topics. As a background to methods used in epidemiology, the statistician should be familiar with linear models for normally distributed observations and with data-intensive methods, including simulation and bootstrapping. Particular courses on procedures applied specifically in epidemiology may be useful, dealing, for example, with analysis of data from prospective studies or case-control designs. They are not essential, however.

Although theoretical foundations are important, critical aspects of the training are more likely to occur in connection with informal techniques used in epidemiological practice. There is hardly any formalised framework which will ensure that students in statistics obtain suitable qualifications in this sense. Thus it is even more important that appropriate training is given in parallel with the formal courses. The ability to translate practical problems into statistical models is an important requirement of this kind. Another one is the ability to select a suitable statistical approach from several available. The student must also see the importance of choosing the right level of complexity in the model building, depending on properties of the data set and the nature of the substantive questions being asked. It is particularly important in epidemiology to take into account potential interactions between risk factors, without creating too complicated models. Furthermore, one must be able to distinguish between minor deviations from a prescribed statistical model, of little practical significance, and major problems which can invalidate the conclusions.

In some situations, the statistician must be prepared to devise a complex statistical strategy including several analytic steps when the choice of method at the later stages depends on preliminary results. Quite frequently in the study of rare diseases, one runs into problems when data sets are too small for ordinary asymptotic methods to apply. Although alternative exact methods may be available, it is necessary to develop an intuitive understanding of the limitations in such data sets. The student must appreciate that there may be several equally valid statistical approaches and should learn to show respect for data-analytic strategies proposed by other researchers. Following the initial data analysis, it may be necessary to go through the observations in different ways, often by elementary statistical methods, to uncover hidden information of potential biological relevance. In other situations, the statistician must be able to carry out a search for interesting patterns in the data.

The training should show how to handle several statistical questions at the same time, not necessarily with formal procedures such as multiple testing, but rather to appreciate that conclusions must be viewed with caution because of the large number of questions being asked. This often applies to joint studies of many risk factors or several related diseases. The student should be well versed in efficient use of graphical methods, both at a preliminary, exploratory stage, and later, especially in the checking of modelling assumptions. More generally, the statistician must know where to search for relevant literature, in statistics or epidemiology, when standard procedures no longer apply. An ability should also be developed to decide whether literature on statistical methodology is worthwhile studying at all to find help in the solution of particular problems. Finally, one should try to create a particular perseverance with practical problems which may at first appear to be unexciting or even hopeless. The ability to analyse 'negative studies' without clear-cut conclusions is also important in epidemiology.

OTHER ACADEMIC SUBJECTS

A sound mathematical foundation is needed, because of the relevance to statistics and also for settling arguments directly related to epidemiology. In addition, it will be useful for the students to see how mathematical methods are applied to completely different areas. Moreover, basic skills in use of computers and programming are required. Students must be exposed to software in common use for general purposes and to statistical software, preferably more than a single major package. Some experience with numerical software will also be of great value.

It has unfortunately been difficult for someone aiming at a degree in statistics to include courses in biological topics. A background in biology may become more essential in the future, as epidemiological studies are expanded to include complex measurements of immediate biological significance, especially in molecular biology. In practice it has been almost impossible to include courses in medical topics, but epidemiological investigations can cover such a wide range of questions that specific medical knowledge can be of little value. An introduction to social sciences may be relevant, but has also been difficult to organise.

REQUIREMENTS AT A PERSONAL LEVEL

A university education may be important in various ways outside the realm of formal curricula. Although university teachers may not always be aware of it, personal characteristics may develop in students during formal training which will be useful in future work. An ability to cooperate with colleagues may be essential, regardless of the path chosen by a graduate in statistics, but it is not easy to foster such personal characteristics through ordinary course work. At least some students should be able to develop an ability to see challenges in problems which are not precisely formulated. This may be particularly important in epidemiological practice. Perhaps students can also develop their own strategies for general problem solving. On the other hand, it may be important to accept that one cannot contribute to the knowledge about risk factors for a disease when the information is too sparse. A related characteristic is the willingness to try entirely different approaches in the handling of practical questions. Some data sets in epidemiology, especially those based on registry data, may appear completely overwhelming in view of the number of variables or observations. The statistician must then be prepared to use particular strategies for extracting relevant information. At the same time, it may be important to develop ones intuition to decide whether particular issues are worth pursuing at all.

A THREE-STEP PROGRAMME FOR FINDING A BALANCED SOLUTION

It is clearly impossible to satisfy all the requirements completely with regard to the training of statisticians who will work in epidemiology. At least in the Norwegian environment, there are not enough resources available to develop a particular degree in biostatistics, let alone a more specialised area such as statistics applied to epidemiology. The training of statisticians must take advantage of courses offered by different departments for other purposes. At the same time, courses in statistics at the basic and intermediate levels must accommodate students with widely divergent interests.

Under such circumstances, with a great many relevant topics and a limited amount of space, it may seem unrealistic to try to adapt courses to the requirements listed above. The problem is certainly not unique to the particular application of statistics considered here but it may be more serious, especially because of all the informal qualifications needed. I shall formulate a programme consisting of three steps for dealing with this general challenge. The programme should be regarded mostly as a vision as it has not been formally implemented, although the underlying ideas are supported by observations made over many years.

Step 1. In the list of particular topics that should be covered, classify the topics into two groups: (A) those which are specifically needed in the education of statisticians in epidemiology, regardless of possible minor variation, and (B) those which play an important role in a general sense only.

The first group A will include many basic mathematical techniques and a considerable amount of material in theoretical and applied statistics. Although teaching methods may vary somewhat even for such topics, the general organisation of the material will often be fixed for all students going through the relevant courses. The second group B includes the remaining material which is considered essential, despite the fact that specific details may not always be important in any particular form. A large part can be regarded as knowledge, ability or insight of a higher order then the usual material included in curricula. This does not imply that students can ignore details in their work with topics in group B, but rather that the objectives will eventually be attained through work with examples or case studies illustrating the general principles. Thus a particular ability can be developed in many different ways. For example, learning a general programming language may be important for later work with complex data sets in epidemiology, but it does not matter very much which language the student becomes familiar with. It is not possible to make an absolute distinction between groups A and B, but it will be relatively easy in most situations to assign each topic to either category. Thus a great many desirable individual skills will belong to group B, although some may also be assigned to group A.

The handling of topics in the second group B may be summarised in this way:

Step 2. For each area in group B representing important topics in a general sense only, specify reasonable minimum requirements which must be satisfied in order to give the students a satisfactory background. Organise courses in such a way that each student will be exposed to every topics to a sufficient extent, at least for the great majority of areas considered.

For example, describe the necessary amount of exposure to real epidemiological data that each student must have. It is not very important exactly what kinds of problems are dealt with or which statistical methods are being used, as long as the problems represent a challenge at a reasonable level of complexity. Similarly, indicate how much experience the students must get in programming, and how much general background in biology is desirable. At this stage, no particular implementations are considered mandatory, although extreme choices of programming languages or bizarre introductions to biology should be avoided. The actual implementation of Step 2 can take different forms, from general recommendations to the students involved, to collaboration with teachers in other subjects and incorporation of related topics into statistics courses. Although the details are not significant, it is still important that Step 2 is carried out according to a definite overall plan.

The broad background given the students through Step 2 should be supplemented by an attempt to combine as much of the different topics as possible:

Step 3. To the extent that it is practically feasible, integrate examples and problems from the second group of topics B in the teaching of specific methodological subjects such as statistics and data analysis in group A.

In some degree, this principle is already being followed in some general statistics courses, but epidemiology has more instructive examples to offer which will be of interest to all statistics students. It is common to use examples from medicine to illustrate logistic regression, but more comprehensive case-studies of applied problems, not related to medicine or biology, will also be important to expose students to practical problems of a general nature. A particular case of the principle described in Step 3 is the writing of a thesis, where it is quite common to use practical problems derived from epidemiology as a basis for considering more theoretical issues.

DISCUSSION

The simple three-step programme described here differs from the policy followed at present especially in the emphasis in Step 2 on topics which are not specified in detail but still play an important role. The present practice has mostly developed without attention to the role that informal statistical skills and non-statistical topics should play in the overall training. The reduction in the amount of theoretical material implied by Step 1 is no major change; neither is the increased emphasis on integration of ideas in Step 3. Step 2 is the real key to educating statisticians who will be better prepared for work in epidemiology.

It is not at all obvious that the recommendations of Step 2 will be possible to carry out within the limits of a 5 year degree. They differ from most policy changes proposed for tertiary education in advocating a greater freedom of choice with regard to the detailed selection of topics. Yet the overall composition of a statistics degree according to this programme is supposed to follow specific general guidelines. It is an essential part of the programme that the more general skills expected of a statistician should be a matter of great concern. This policy contrasts with that recommended in Step 1 for topics which are needed in specific detail. When group A has been cut down to a reasonable minimum, mostly including mathematical and statistical topics, it is hard to see that students can be allowed very much choice in this category.

The three-step programme aims at maintaining standards in methodological subjects and minimising the gap between university education and work in epidemiology. It can fail in two ways: the graduates may not reach the theoretical and methodological standards in group A required by future work, or the time spent on topic in group B may be largely irrelevant to the real problems encountered later. The first situation should not occur if the recommendations in Step 1 are followed accurately. I believe the second situation is unlikely to occur if enough care is taken in the implementation of Step 2, supplemented by Step 3. The final proof that a programme of this kind can produce an appropriate balance between theory and practical training must of course wait until it has been put into practice. Even then, it may take many years before the situation can be assessed on the basis of the work experience of graduates. A failure might simply mean that the statisticians working in epidemiology are being replaced by other professionals. The programme proposed represents a compromise between conflicting demands, but recognising that there are potential conflicts and trying to achieve a reasonable balance, seems the best policy.

Is epidemiology essentially different from other areas that statistics graduates move into? Corresponding programmes for most areas would probably have less extensive groups B of topics of general importance, and carrying out Step 2 would not have the same significance. As work in the insurance industry is at present the major competitor on the job market for statistics graduates in Norway, it is natural to compare with the specialised courses given in insurance mathematics at Norwegian universities. My suggestions for epidemiology or general medical applications are very modest compared to the adaptations already made to the needs of insurance. Of course, many of the recommendations given here are also relevant for more general work in statistical consulting. It is has been suggested that training in applied epidemiology should be based on the 'philosophy of learning while doing' (Thacker and Buffington, 2001), but as so many of the informal skills needed are related to use of statistical methods, it seems essential that statisticians entering the area should be prepared in a different way. To develop ones intuition in connection with data analysis takes a long time.

The classification of skills needed by a statistician is in a general sense similar to that used by Bryce et al. (2001) in connection with the curriculum guidelines for undergraduate degrees in statistical science approved by the American Statistical Association. A major difference lies in the level of professional work aimed at. With a specific area of application in mind, it also becomes much easier to recommend a background in other subjects. Schlesselman (1996) underlined how important it is for biostatisticians working epidemiology to have a basic biological insight. Pocock (1995) also pointed out that the training of statisticians who will work in medical applications has mainly been far too theoretical. Previous discussions have not offered many specific solutions, however, except for emphasising the need for change. Apparently the question of finding a reasonable balance between theoretical and informal topics has not been addressed in the same way. This may be related to the general problem of formulating what the basic elements of applied statistical practice and statistical thinking actually are (Pfannkuch and Wild, 2000), an issue which is likely to become more important to teachers of statistics at all levels in the future. Further developments in the application of biostatistical methods to epidemiology may in particular take into account special aspects of the data collection (Levy and Stolte, 2000) and informal aspects of data analysis may then play an even greater role.

REFERENCES

Bryce, G. R., Gould, R., Notz, W. I. and Peck, R. L. (2001). Curriculum guidelines for bachelor of science degrees in statistical science. *The American Statistician, 55,* 7-13.

Buffington, J., Lyerla, R. L. and Thacker, S. B. (1999). Nonmedical doctoral-level scientists in the Centers for Disease Control and Prevention's Epidemic Intelligence Service, 1964-1997. *American Journal of Preventive Medicine, 16,* 341-346.

Ellenberg, J. H., Gail, M. H. and Geller, N. L. (1997). Conversations with NIH statisticians: interviews with the pioneers of biostatistics at the United States National Institutes of Health. *Statistical Science, 12,* 77-81.

Levy, P. S. and Stolte, K. (2000). Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Statistical Methods in Medical Research, 9,* 41-55.

Pfannkuch, M. and Wild, C. J. (2000). Statistical thinking and statistical practice: themes gleaned from professional statisticians. *Statistical Science, 15,* 132-152.

Pocock, S. J. (1995). Life as an academic medical statistician and how to survive it. *Statistics in Medicine, 14,* 209-222.

Schlesselman, J. J. (1996). Biostatistics in epidemiology: a view from the faultline. *Journal of Clinical Epidemiology, 49,* 627-629.

Thacker, S. B. and Buffington, J. (2001). Applied epidemiology for the 21st century. *International Journal of Epidemiology, 30,* 320-325.