Beyond the Confidence Interval

CHARLES POOLE, MPH

"It is essentially consideration of intellectual economy that makes a pure significance test of interest."[1]

Introduction

Until recently, the discussion of random error and how to account for it in epidemiologic research has resembled the sound of one hand clapping. Writers on the subject have decried the convention of significance testing as arbitrary, uninformative, and conducive to misinterpretation[2–4]; but defenses of the practice have been rare and, for the most part, halfhearted. Meanwhile, significance testing has been unperturbed as the mainstay of statistical analysis in epidemiology, despite a superficial shift by many investigators from the reporting of p-values to the reporting of confidence intervals.

Joseph Fleiss has now risen to defend significance testing against the criticisms that have thus far gone largely unanswered.[5–7] Fleiss's argument goes essentially as follows: We need to make decisions in science, especially in a science as closely linked to personal and public policy choice as epidemiology. Epidemiologists, as applied scientists, need to agree by consensus on pre-specified criteria so that the bases for their decisions will be explicit. The convention of significance testing, though not without limitations and potential for abuse, serves epidemiologists well as a reasonable means of facilitating scientific decision making.

Alexander Walker[8,9] and others[10–13] have expressed arguments in contraposition to Fleiss's. The confidence interval, these advocates claim, provides more infomation about random error than does the p-value or the significance test. Confidence intervals focus one's attention on the magnitude of an estimate of a meaningful parameter (e.g., the incidence rate ratio) and, as a separate matter, on the precision of that estimate. Significance tests, on the other hand, blend together the magnitude of the estimate and the hypothetical role that random error may have had in producing it. Thus, despite whatever virtue they may have as decision-making tools, significance tests are inferior to confidence intervals as conveyances of information.

I wish to propose another perspective of this controversy. It is not a new perspective, but neither has it received the critical consideration it deserves. It is a standpoint from which the distinction between significance testing and interval estimation fades to virtual irrelevance. I offer it in two ways: first, in the spirit of criticizing both significance testing and the currently popular interpretation of confidence intervals; and second, as an affirmative argument in support of the full explicitness for which Fleiss admirably calls but of which both testing and interval estimation fall miserably short.

Premises

Although science and decision making are both important, they are not the same. In science we seek to learn, to explain and to understand. In decision making, we seek reasons to act or to refrain from acting. It demeans neither of these enterprises to acknowledge that they are different from each other. Consider on the one hand the theory that contraceptive diaphragms cause urinary tract infections[14] and, on the other, decisions about the use of diaphragms. The causal theory is either true or false. We can criticize and empirically test it, but we cannot *decide* that it is or is not true. Its truth or falsity is completely independent of any decision we might make about it. What we *can* decide is what our actions, both personal and public, will be with respect to the use of diaphragms.

Clinicians, health policy makers, and citizens practice decision making all the time. They recognize that there is much more to every single decision in medicine and public health than the status of the critical discussion of a causal theory, let alone the contribution of a single study to that discussion. In collaboration with Stephan Lanes[15] and Kenneth Rothman,[16] I have made the argument that these other considerations create a distinction between epidemiologic science and public health policy choice. In a culture that reveres scientists and scorns politicians and bureaucrats, there may be a temptation to infer that we mean to denigrate policy making by separating it from science. To the contrary, our hope is for the legitimacy and importance of policy analysis to become apparent.

Our distinction between the conduct of health-related science and the making of health-related decisions is far from original. It may be viewed as but an illustration of Karl Popper's demarcation between science and non-science.[17–19] Commensurate notions have been expressed in statistics, as in John Tukey's distinction between decisions and conclusions.[20] In at least one school of public health, biostatistics students may choose to concentrate in biostatistics *or* in "health decision sciences."[21] Among the most illuminating views on this subject from the statistical perspective are the arguments made by the statistical theoreticians D. R. Cox and D. V. Hinkley.

In the preface to their text, *Theoretical Statistics*,[1] Cox and Hinkley draw the sharpest distinction between "the theory of statistical methods for the interpretation of scientific and technological data" and statistical decision theory, "where statistical data are used for more or less mechanical decision making." The unmistakable rationale for this distinction is that scientific research is our empirical way of trying to increase our explanatory understanding of the world and not our way of going about the task of living in the world. Although our ever-tentative understanding has a bearing on our decisions, attempting to understand and deciding to act are not identical thought processes. Cox and Hinkley recognize this difference when they contrast statistical problems in science with statistical decision problems. With respect to decisions, they state that "the analysis has the aim not of, in some sense, assessing the information that is available about the unknown parameter but rather that of choosing between a number of clearly formulated alternative courses of action."[1]

Fleiss's call for pre-specified decision rules[5–7] seems to belong more to the statistics of decisions than to the statistics

Address reprint requests to Charles Poole, Epidemiology Resources Inc., 826 Boylston Street, Chestnut Hill, MA 02167. He is also with the Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.

of science. Yet Cox and Hinkley assert that, *even within decision analysis,* "it is unlikely that single major decisions will or should be made by mechanical application of a decision rule. The better approach will be to isolate for critical discussion the separate aspects entering into the final decision."[1] Popper, of course, has argued that critical discussion is the hallmark of scientific evaluation.[17–19]

Critical discussion requires thought and is therefore difficult. "Mechanical application of a decision rule" requires no thought and is therefore easy (once the decision rules have been set). The naive hope that we can find an easy way out of critical discussion in science is thus a proposal to abandon our ability and duty to think. Cox and Hinkley, in the prefatory quotation I chose for this essay, found an elegantly polite expression for the abandonment of thought: "intellectual economy." There can be no more cynical or pessimistic view of science than the suggestion that scientists should economize their intellects.

Criticism of Significance Testing

When in Chapter 7 of *Theoretical Statistics*[1] Cox and Hinkley leave significance testing and come to the topic of interval estimation, they state that they have finally arrived at "the central problem of statistical inference." As if to substantiate this claim, Cox published in 1977 a paper, "The Role of Significance Tests,"[22] in which he came to the following conclusion:

> The central point is that statistical significance is quite different from scientific significance and that therefore estimation, at least roughly, of the magnitude of effects is in general essential regardless of whether statistically significant departure from the null hypothesis is achieved. It is only when the qualitative result of such estimation is clear from the context that the result of the significance test stands almost on its own as the main summary of the analysis.[22]

This appraisal stands in stark contradistinction to Fleiss's claim that our first priority in the analysis of epidemiologic data should be to "establish the reality" of an association by the means of testing its statistical significance.[5]

Fortunately, one does not need to be a theoretical statistician to appreciate the extreme "intellectual economy" that inescapably accompanies significance testing. The practice seems to be promoted on the theory that scientists need to protect themselves and others against the dangers of thinking. After all, as Fleiss notes with alarm, if we are allowed to think we might arrive at different interpretations. He thus claims that we need safeguards against diversity of interpretation, against imaginative theorization, and against the possibility that "my substantive difference may be your trivial difference."[5,6]

To the contrary, the notion that it is hazardous for scientists to think is itself an exceedingly dangerous myth. A more reasonable proposal is that we need to encourage thinking in science in order to safeguard ourselves and others from the hazards of rituals like significance testing. I shall illustrate with an example, selected from the history of occupational epidemiology, in which a single confidence interval would have made all the difference in the world.

During the decade between 1968 and 1977, there was only one epidemiologic study worthy of note on an important occupational health topic: the role of cigarette smoking in mediating the effect of asbestos exposure on the occurrence of lung cancer. This study[23] reported a very strong association between asbestos exposure and lung cancer among cigarette smokers. Among the non-smoking asbestos workers, on the other hand, no lung cancer deaths were observed; *but only 0.05 such deaths were expected.*

For their part, the authors[23] were unimpressed with the result for non-smokers. Other epidemiologic observers were less cautious, however. Referring to this study alone, the following interpretations appeared: "We conclude from the epidemiological findings that asbestos induces mesothelioma of the pleura and peritoneum, but not by itself [cancer] of the bronchus"[24] and "[A]pparently, asbestos will produce lung cancer only in smokers."[25] This ill-begotten message found its way into health education materials for employees in the asbestos industry. One pamphlet proclaimed, "Studies show that if you don't smoke cigarettes, asbestos does not increase your risk of lung cancer."[26]

I computed (with the mid-p method) a 90 per cent confidence interval for the estimated rate ratio of zero for the non-smoking asbestos workers in this study. The limits of this interval are rate ratios of zero and 46.1. The interpretations quoted above would have been difficult to issue in the face of such a confidence interval. No test of statistical significance, alone or accompanied by a statement of statistical power, would have conveyed the extreme imprecision of this study's data on non-smokers nearly as well as a confidence interval would have.

Criticism of the Popular Interpretation of Confidence Intervals

My criticism in this section is not of confidence intervals themselves, but of the way they are commonly interpreted. In epidemiology today, confidence intervals are usually taken as nothing more than tests of statistical significance. The way this interpretation proceeds is by looking to see whether the null value of the parameter is inside the interval or outside of it. It takes no thought to accomplish this task; a computer could do it easily (and not a very big computer at that). Curiously, however, it is a little bit harder than comparing a p-value to an alpha-level, especially if the data lie on the so-called "threshold of significance." And certainly, confidence intervals take up more room in tables and text than p-values, asterisks, or the abbreviations "S" and "NS".

Why have epidemiologists changed their form of presentation to a more awkward one, but not changed their (aversion to) thinking? This would be a good question for a sociologist of science to investigate. I have no theory to test, nor am I particularly interested in developing one. What interests me is the amount of useful information and critical discussion in the reports of epidemiologic research I read. Tests of significance, whether obtained by comparing p to alpha or by looking for null values within confidence intervals, are worse than useless. They are misleading. Consequently, they inhibit critical discussion.

W. Douglas Thompson comes close to this same conclusion when he writes that "it is important to recognize that population values just beyond the confidence limits are only slightly less likely to have given rise to the observed data than are some of the values included in the interval."[12] But in his proposal to use confidence intervals to separate parameter values categorically into "those values with which the observed data are compatible and those with which they are incompatible," he adopts an outlook that differs little from the dichotomization of results into those that are and are not "significant".

The following example illustrates how similar (and how similarly misleading) Fleiss's approach to significance testing

TABLE 1—Prevalence Ratio Estimates from a Study of Spermicides and Down Syndrome[27]

Control Group	Prevalence Ratio	
	Point Estimate	90% Confidence Interval
Random Controls	3.6	1.2–9.0
Congenital Heart Disease Controls	2.8	0.9–7.3

and Thompson's approach to the interpretation of confidence intervals can be. The example is a recent case-control study of parental spermicide use in relation to the prevalence of Down syndrome at birth.[27] The author of this study suspected recall bias, so in addition to a random sample of newborns he used a second control group of children born with congenital heart disease. The results are summarized in Table 1.

The author made no claims about the presence or absence of statistical significance in these data. Nevertheless, a set of observers subsequently wrote that "spermicide use was significantly more common only when Down's cases were compared with normal control subjects, not other malformed infants. This suggests that recall bias may have caused the higher rate of spermicide use reported in the Down's group."[28]

It is important to note that this interpretation follows the advice of both Fleiss and Thompson almost to the letter. It cannot, therefore, be viewed as an "abuse" of either author's recommended interpretive procedure. That is to say, the observers interpreted the confidence intervals as significance tests and interpreted the significance tests as "criteria for inferring that an association is real."[6] Equivalently stated, the observers took the confidence intervals as "zones of compatibility" with the data and took parameter values outside the intervals to be incompatible with the data.[12] In the next section, I shall show a simple method of graphical presentation that would have put these data in a considerably different interpretive light.

In Pursuit of Explicitness

Although I disagree with the notion that science is decision making, I agree fully with Fleiss's call for epidemiologists to be more explicit. According to the dictionary on my desk, to be explicit is to be "free from all vagueness and ambiguity," to be "fully developed and formulated," and to be "unreserved and unambiguous in expression." I doubt that anyone would refuse to endorse these attributes in reports of epidemiologic research. How closely do significance tests and interval estimates, when used as surrogate tests, approximate them?

The answer is, "Not very well at all." The reason is that the selection of the level of significance or confidence is arbitrary. Fleiss wants all thoughts as well as procedures in epidemiology to be reproducible,[7] yet he describes with pride the long deliberations in which he and his colleagues engaged in their struggle to select a level of significance for a certain analysis.[5] Could other investigators reproduce these deliberations, even if they wanted to? I think not. Therefore, this example of significance testing in practice is an illustration of an inexplicit, irreproducible process.

Fortunately, there is a simple and informative means to avoid vagueness, ambiguity, incomplete development, and partial formulation of the deductions to be drawn from our models of random error. The solution is to present and

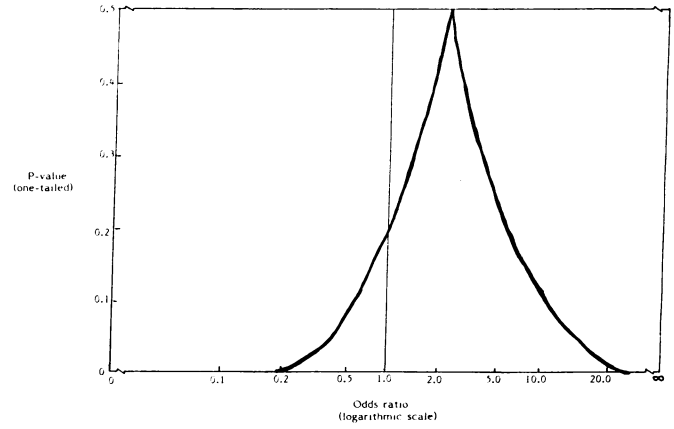

FIGURE 1—P-value Function Corresponding to Thompson's Figure 1 (reference 12)

interpret the graph of the p-value function in its totality.[cf.29,30] The p-value function may be considered the graph of all possible p-values (the p-value to which we normally refer being simply the *null* p-value) or the graph of all possible confidence limits. There is an analogous graph in the likelihood function.

Figure 1 presents a sketch of all possible confidence limits for the data represented by Thompson's Figure 1.[12] (The x-axis is on a logarithmic scale in order to symmetrize the range of the rate ratio measure.) This graph tells us the degree of compatibility with the data that can be deduced *for every parameter value* from the selected family of probability models. I have presented this and subsequent graphs just as I sketched them to show how easy they are to construct from point estimates, confidence limits and p-values from published papers. For orientation, the reader can find the 95 per cent confidence limits of 0.4 and 17.3 on this graph at the points corresponding to p = 0.025. The point at which the curve crosses the vertical line representing a rate ratio of 1.0 corresponds to the (null) p-value that would be compared to alpha in a (one-tailed) significance test. The point estimate at the top of the curve may be thought of as a zero per cent confidence interval.

Graphs such as these are much more explicit than significance tests and much more informative than confidence intervals. Technically, the graphs are infinitely more explicit and informative because the number of possible significance or confidence levels is infinite. The graphs completely avoid the arbitrariness of selecting a single significance or confidence level from this infinite menu.

Is it possible, as Walker claims,[8] to have too much information? I think not, as long as the information is summarized in a useful form as in these graphs. Their full utility comes, of course, when one result is compared with another. Figure 2 shows the p-value function for the data represented by Thompson's Figure 2. The contrast with Figure 1 is sharp. The two point estimates are virtually identical, but one estimate is much more precise than the other.

Figure 3 shows the p-value functions for the two estimates from the illustrative study of spermicides and Down syndrome.[27] The magnitude and precision of the estimates are virtually the same from the two control groups. The data thus tend to refute, not support, the theory that recall bias was responsible for the departure of the estimate using random controls from the null value of one. The interpreta-

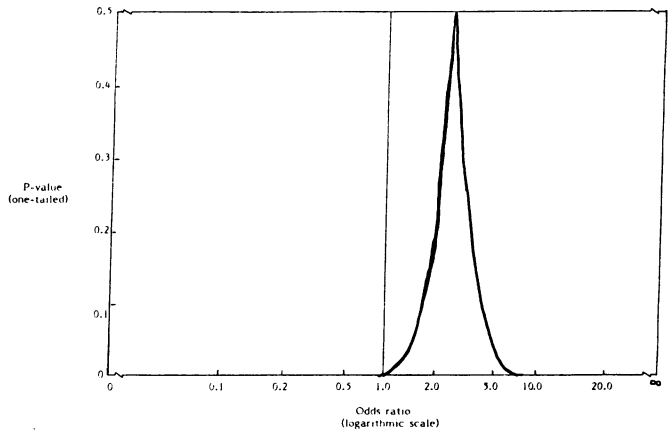

FIGURE 2—P-value Function Corresponding to Thompson's Figure 2 (reference 12)

tion offered by the observers[28] would not have been possible if they had looked at graphs like these; yet, this misguided inference clearly was possible in the light of the published confidence intervals.

Conclusion

Significance testing is deeply ingrained in the epidemiologic consciousness. Consider the following experience, which I offer in the spirit of balancing Fleiss's complaints about editors who "insidiously" discourage significance testing.[5] Lanes and several colleagues, including myself, submitted a paper in which we presented rate ratio estimates from a case-control study on possible etiologic relations that had so far been largely overlooked. The study was small and the estimates were imprecise, but we considered them worthy of consideration. We emphasized the imprecision of the estimates by drawing attention to the width of the 90 per cent confidence intervals we had constructed. We made no claim of "reality" for the theoretical relations, no conclusion of causality, and certainly no assertion about statistical significance.

In the course of informing us that our paper had been rejected (it has since been accepted elsewhere), the editor referred in particular to the comments of one reviewer. The reviewer had somehow transferred his or her own interest in statistical significance to us by falsely accusing us of "claiming statistical significance" for one of the results and of claiming "near-significance" for another. The reviewer further accused us of "switching" from 95 per cent to 90 per cent intervals so that we could make these alleged claims. This example shows the kind of thoughtful criticism that significance testing inspires.

I hope that an occasional researcher will have the temerity to try to publish a complete p-value or likelihood function for the main result from an epidemiologic study. Short of such bravery, I ask only that investigators and readers sketch such graphs every now and then in privacy, just to remind themselves of the shape of the function whose image should be evoked by p-values, point estimates, and confidence limits.

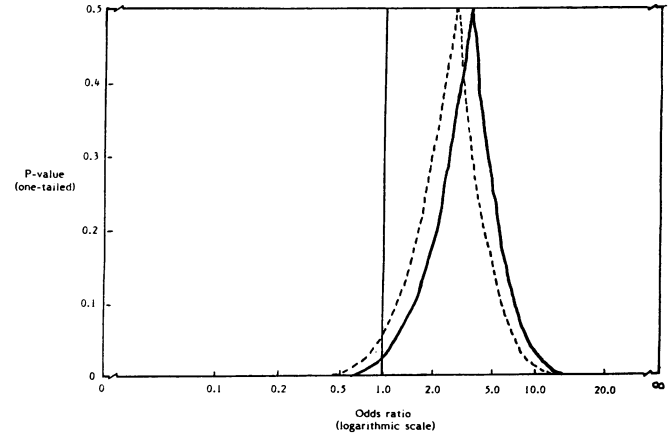

FIGURE 3—P-value Functions Corresponding to the Estimates in Table 1 with Random Controls (solid line) and Congenital Heart Disease Controls (broken line)

ACKNOWLEDGMENTS

This essay is based on a paper delivered at the nineteenth annual meeting of the Society for Epidemiologic Research in Pittsburgh, PA on June 19, 1986.[31]

I thank Sander Greenland for criticism and literature citations, and Olli Miettinen for assigning a p-value function to be drawn as homework in a course at the Harvard School of Public Health.

REFERENCES

1. Cox DR, Hinkley DV: Theoretical Statistics. London: Chapman and Hall, 1974.
2. Rothman K: A show of confidence. N Engl J Med 1978; 299:1362–1363.
3. Salsburg D: The religion of statistics as practiced in medical journals. Am Statis 1985; 39:220–223.
4. Gardner MJ, Altman DG: Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986; 292:746–750.
5. Fleiss JL: Significance tests have a role in epidemiologic research: reactions to A. M. Walker. (Different Views) Am J Public Health 1986; 76:559–560.
6. Fleiss JL: Confidence intervals vs significance tests: quantitative interpretation. (Letter) Am J Public Health 1986; 76:587.
7. Fleiss JL: Dr. Fleiss responds. (Letter) Am J Public Health 1986; 76:1033–1044.
8. Walker AM: Reporting the results of epidemiologic studies. Am J Public Health (Different Views) 1986; 76:556–558.
9. Walker AM: Significance tests [sic] represent consensus and standard practice. (Letter) Am J Public Health 1986; 76:1033. (See also Journal erratum 1986; 76:1087.)
10. Foxman B, Frerichs RR: Response from Drs. Foxman and Frerichs. (Letter) Am J Public Health 1986; 76:587.
11. Rothman KJ, Yankauer A: Editors' note. Am J Public Health 1986; 76:587–588.
12. Thompson WD: Statistical criteria in the interpretation of epidemiologic data. (Different Views) Am J Public Health 1987; 77:000–000.
13. Savitz D: Comments received on significance tests and confidence intervals. (Letter) Am J Public Health 1987; 77:000–000.
14. Foxman B, Frerichs RR: Epidemiology of urinary tract infection: I. diaphragm use and sexual intercourse. Am J Public Health 1985; 75:1308–1313.
15. Lanes SF, Poole C: "Truth in packaging?" The unwrapping of epidemiologic research. JOM 1984; 26:571–574.
16. Rothman KJ, Poole C: Science and policy making. (Editorial) Am J Public Health 1985;75:340–341.
17. Popper KR: The Logic of Scientific Discovery. 2nd Ed. New York: Harper & Row, 1968.
18. Popper KR: Conjectures and Refutations: The Growth of Scientific Knowledge. New York: Harper & Row, 1968.
19. Popper KR: Objective Knowledge: An Evolutionary Approach. Rev. ed. Oxford: Clarendon Press, 1983.
20. Tukey JW: Conclusions v. decisions. Technometrics 1960; 2:423–433.
21. Official Registry of Harvard University: Harvard School of Public Health 1986; 7:14–15.
22. Cox DR: The role of significance tests. Scand J Statist 1977; 4:49–70.
23. Selikoff IJ, Hammond EC, Churg J: Asbestos exposure, smoking and neoplasia. JAMA 1968; 204:106–112.
24. Hoffmann D, Wynder EL: Smoking and occupational cancers. Prev Med 1976; 5:245–261.
25. Cole P, Goldman MB: Occupation. *In:* Fraumeni JF Jr (ed): Persons at High Risk of Cancer: An Approach to Cancer Etiology and Control. New York: Academic Press, 1975, 167–183.

26. W. R. Grace, Construction Products Division: There are some things you should know. Employee pamphlet.
27. Rothman KJ: Spermicide use and Down's syndrome. Am J Public Health 1982; 72:399–401.
28. Mills JL, Reed GF, Nugent RP, et al: Are there adverse effects of periconceptional spermicide use? Fertil Steril 1986;43:442–446.
29. Folks JF: Ideas of Statistics. New York: John Wiley & Sons, 1981, 182–183.
30. Miettinen OS: Theoretical Epidemiology: Principles of Occurrence Research in Medicine. New York: John Wiley & Sons, 1985, 107–128.
31. Poole C: Beyond the confidence interval (Abstract). Am J Epidemiol 1986; 124:527.

WHO Global Program for Appropriate Health Care Technology

The World Health Organization (WHO) has now formalized a global program to address the issue of the appropriate use of technology in health care, defined as drugs, devices, equipment and organization within the health care system.

The Global program meshes into WHO's worldwide *Health for All* strategy and its targets. These propose that by the late 1980s each Member State should have assured quality of services and worked out its technology assessment needs, and by the late 1990s should have or have access to an operational mechanism for systematic monitoring and evaluation.

The program operates through a network of institutions and resource people working on common projects dealing with specific problems.

The aims of this program are to identify the technologies in need of assessment, whether already in use, under development or forecast; to develop methods for assessing technology by studies and literature review; to convene consensus groups; to analyze, validate and disseminate information; to publish reports with global implications; to establish contact with governments, intergovernmental and nongovernmental organizations, industry and consumer groups; to develop national models to advise on standards for quality assurance; to promote educational efforts aimed at providing health workers, policy makers and the public with a proper understanding of health care technology and the problems of its transfer, and to promote contacts with the mass media as well as specialized journals, media and industry.

In collaboration with Member States, insurance companies, health professionals and consumer groups, the following key issues have been identified as high priority:

- Communication Technologies: Information, computers in health care, health policy and management
- Comparison of Variation in Health Care Practice
- Budgetary Incentives and Disincentives for Appropriate Use of Technologies
- Assessment and Use of Medical Technologies: Methodology, e.g. insulin pump study
- Laboratory Technologies
- Imaging Technologies. Basic Radiological Systems (BRS) and Magnetic Resonance Imaging (MRI)
- Perinatal Technologies: e.g. ultrasound
- Safety in Health Care: Hospital infection control, biosafety, prevention of allergy, oral health
- Drug Utilization: Antibiotics, iron and respiratory infection.

For further information, contact: Manager, Global Program for Appropriate Health Care Technology, WHO, Nyropsgade 18, DK-1602, Copenhagen-V, Denmark.