3.

EFFICACY AND SAFETY ASSESSMENT: HISTORY AND CASE STUDIES

3.

EFFICACY AND SAFETY ASSESSMENT: HISTORY AND CASE STUDIES

The purpose of this chapter is to provide an initial perspective on the development and current state of issues relating to efficacy and safety. To provide this perspective, the chapter briefly traces the evolution of interest in estimating the efficacy and safety of medical technologies. Seventeen brief case studies are presented to illustrate many of the issues relating to such estimation.

EVOLUTION OF INTEREST IN EFFICACY AND SAFETY ESTIMATION

> Taking the expression *clinical trial* in its widest possible sense—that is, to cover the test of any therapeutic procedure applied to a sick person—it is obvious that the clinical trial must be as old as medicine itself. Even the witch-doctor trying out for the first time a new and nauseating compound must surely, like Alice nibbling at the mushroom in Wonderland, have murmured to himself 'which way?' —though he would no doubt have concealed his anxiety from his patient with the customary bedside manner. Such personal observations of a handful of patients, acutely made and accurately recorded by the masters of clinical medicine, have been, and will continue to be, fundamental to the progress of medicine (165).

As Bradford Hill's comment above indicates, the development of statistical techniques for evaluating efficacy and safety does not lessen the-historical importance of clinical judgment and individual decisionmaking: modern evaluation techniques should complement the traditional,

Today's techniques and the willingness to use them did not come about overnight; nor did they come about because physicians today are more concerned than their predecessors about the outcomes of medical practice.

As early as the 18th century, statistics and probability techniques were used, though rarely, in support of medicine and public health. Cotton Mather, the American clergyman, reported in 1721 that in the Boston smallpox epidemic of that year more than 1 in 6 persons who were not inoculated against the disease died, but that only about 1 in 60 who were inoculated did so. Though his mathematics were crude by today's standards and his "experimental design" certainly weak, this effort represents one of the very early statistical tests of the benefit of a medical technology (316). In 1759, Benjamin Franklin published an account of the success of vaccination in Boston (122). His report* contains mathematical analyses of the results of vaccination, but it is also an early example of a medical "review article" and (in many respects) of a policy analysis.

*"Some account of the success of inoculation for the smallpox in England and America. Together with plain instructions, by which any person may be enabled to perform the operation, and conduct the patient through the distemper. "

In later years of that century, mathematicians, such as Bernoulli, studied the efficacy of various vaccines. The relatively widespread introduction of vaccination after 1800, with the associated dramatic and clearly observable improvements in mortality and morbidity from the target diseases, greatly diminished the interest in statistical studies of inoculation and its relation to the prevention of disease.

Toward the end of the 18th century, though, interest arose in the use of statistics to study the effectiveness of treatments. For example, in **1793,** when Benjamin Rush announced that he had discovered a definite cure for yellow fever, William Cobbett, an English politician visiting Philadelphia, inquired as to Rush's proof that the treatment, largely bleeding and purging, was effective. Rush, like most physicians of the time, did not keep complete case records, so Cobbett assembled statistics from "bills of mortality" and discovered a positive correlation between the numbers being treated and mortality rates. Though his calculations omitted many important variables, Cobbett's claim of harm rather than cure did seem to have some basis. As Shryock states:

> Here, at any rate, was an appeal to statistical evidence against a particular therapeutic procedure—a rather unique appeal for the period. So unique, was it, indeed, that it received small attention from either doctors or laymen. Cobbett was eventually convicted of slander, fined, and practically driven out of town. Yet he actually suggested the use of statistics in therapeutic research. What one man saw, in the heat of controversy, others would realize sooner or later in the course of calm investigation(316).

In 1810, Laplace's classic study of the calculus of probabilities included a strong statement of the potential of that technique in medical research. By the middle of the 19th century a trend toward increased interest in the use of statistics in medicine became evident (203). Paralleling this interest was the increasing reliance of medicine on scientific methods and on discoveries of the natural sciences. By combining the new emphasis on linking symptomatology to treatment to pathology with the techniques of mathematics, Pierre Louis of France was able to study the effectiveness of various therapies.

His statistical techniques were simply a sophisticated method of extending the experience and quantifying the impressions of physicians (316). These techniques put forth by Louis and others in the 1800's were resisted, sometimes for logical reasons, but rapidly became an integral part of clinical investigation.

Because so many of the therapies in vogue in the first half of the 19th century were not efficacious, the increasing use of statistical techniques began to reduce substantially the number of accepted treatments. This "therapeutic nihilism" was not countered by the development of efficacious therapies to replace those discredited.

Thus, paradoxically, the advance in medical science represented by submitting therapies to quantitative evaluation was one of the contributions to the fairly widespread loss of public confidence in medicine during the last half of the 19th century (316). In this period, however, the study of microorganisms and their role in disease was beginning to produce a base for later prevention and treatment.

These and other developments in medical science led to a number of striking advances, beginning roughly at the turn of the century. Mortality and morbidity declined rapidly—perhaps substantially as a result of improvements in the environment and personal habits, but also often because of medical preventive and therapeutic measures. The use of statistics and the scientific assessment of efficacy and safety grew slowly and were not generally regarded as a critical aspect of medicine. The reasons were numerous: the successes of the first third of the 20th century seemed evident, public confidence in medicine was high, and few legal requirements to demonstrate efficacy and safety existed.

Increasingly, however, medicine has been directed toward the chronic and degenerative diseases. Because it is more difficult to study the effects of treatments for such conditions, the state of medicine's ability to evaluate efficacy and safety became more critical. This situation, added to the growing interest of a small number of individuals (such as Bradford Hill) and improvements in the mathematical techniques available, led to debate as to the appropriate form, role, and magnitude of efficacy and safety evaluation. Miller states:

> The conquest of most of the acute and chronic infections in the developed world has left medicine now preoccupied with a large number of diseases of multiple etiology and long duration, where the assessment of therapeutic results presents real difficulties(242).

Barnes agrees th0 - scientific studies are essential, but he contrasts the relatively unsophisticated techniques of the early 20th century to today's techniques:

> Possibly the most critical and central defect in these cited studies of (late 19th, early 20th century) innovative surgical therapy is the lack of control experience, The concept of controls appeared to be totally unknown to the surgeons of this period. . . . (23)

The reluctance of physicians to embrace a statistical approach to effectiveness continued into the 20th century (163). In 1921, a writer in the *Lancet* asked whether the quantitative method was an "important stage in the development of (medicine)" or a "trivial and time-wasting ingenuity as some hold" (164). By 1971, Hill was able to report the medical community's answer, which was a "remarkable and increasing acceptance of the method. " In 1938, the Federal Food, Drug, and Cosmetic Act was passed, requiring th0t the safety of new drugs be demonstrated by scientific investigation before marketing was allowed. Cochrane believes th0t a "critical step forward" in the use of experimental methods in clinical medicine took place in 1952, when Daniels and Hill published their study of the efficacy of chemotherapy for pulmonary tuberculosis (72). The 1962 Kefauver-Harris Amendments to the Federal Food, Drug, and Cosmetic Act added the requirement th0 - efficacy as well as safety be demonstrated for drugs. Since 1976, certain medical devices have been required to be demonstrated as safe and effective.

The 1970's present a contrast. The techniques available to estimate efficacy and safety are more sophisticated th0n ever, and 0 - the same time concern is increasing about too little, too *much*, and *inappropriately timed* evaluation of efficacy and safety. These issues are discussed in chapter 6, but the next section of this chapter illustrates many of them by presenting 17 brief case studies.

CASES ILLUSTRATING EFFICACY AND SAFETY ISSUES

As mentioned above, medical technology has transformed medical practice in the past several decades by making new preventive, diagnostic, and therapeutic tools available to the medical care system. On the other hand, the accelerating pace of technological development has raised a number of troubling issues. Questions are being raised about whether current research and development efforts are directed at developing the most desirable technologies, whether new technologies are adequately assessed for safety and efficacy before they come into widespread use, and whether valuable technologies come into general use as rapidly as they might.

One way to address these issues is to assess the efficacy and safety of new medical technologies prior to diffusion and, when possible, existing medical technologies where serious doubt exists as to their effects. The nature, aims, current status, cost, and policy

implications of medical technologies will all influence evaluation of their safety and efficacy. The 17 case studies presented here illustrate a variety of points concerning the efficacy and safety of medical technologies. They do not, however, illustrate all the possible issues or concerns, nor are they intended to be complete reviews of the efficacy and safety of the particular technologies used. The cases may sometimes touch on effectiveness or cost-effectiveness. Although this report is not concerned directly with these issues, the relationships between efficacy, effectiveness, and cost-effectiveness are important. The cases can often serve as an introduction to the interworkings of these concepts.

Cases of accepted efficacy are included, as are cases of uncertain efficacy. Expensive technologies are represented, as well as relatively inexpensive ones (at least on a per-unit basis). Some of the technologies have already been widely diffused; others are only beginning to diffuse. Several cases show considerable Federal Government involvement; others, relatively little. Taken together, they are intended to demonstrate some of the complexities that must be recognized if medical technologies are to be evaluated for efficacy and safety.

Case 1: Pap Smear for Cervical Cancer*

The Pap smear test is an analysis of cells taken from the uterine cervix (neck of the uterus) to screen for cancer of the cervix. These cells are usually and most effectively obtained by scraping the cervix. The smear may be taken in a doctor's office, clinic, or hospital. The procedure is quick and simple but may cause some discomfort. Its safety has never been questioned.

In 1973, cervical cancer accounted for 6,000 deaths and ranked fifth among cancers for women as a cause of death (U.S. Vital Statistics, 1975). The death rate from, and incidence of, cancer of the cervix have been declining in the United States since before screening began. The American Cancer Society (ACS) estimates 20,000 new cases annually. The disease is more prevalent among women of lower socioeconomic classes, women who begin sexual intercourse at an early age, and women who have many sexual partners.

The Pap smear has been widely promoted for annual use in the United States. In 1973, 75 percent of U.S. women over the age of 17 had had a Pap smear at least once and nearly half had had one in the year prior to the survey (National Center for Health Statistics, 1975). No other country in the world has achieved this level of screening.

The average cost of examining a Pap smear by a cytological laboratory is about $5, with a range of $3 to $10. The actual costs of screening are, in fact, higher, as they should include the cost of the gynecologist or clinic unless the visit is for other purposes. One must also count the costs of followup for definitive diagnosis for those women who have abnormal Pap smear results but do not have any disease (false positives).

The test was generally accepted when it was introduced in 1943. In recent years, however, health professionals, particularly epidemiologists, have disagreed over the efficacy of the Pap smear as a screening device. The controversy has centered on three issues: the natural course of the disease, the accuracy of the test, and the efficacy of screening in lowering cervical cancer mortality rates.

The test results of the Pap smear are usually reported in five classes: I—normal, II—atypical, III—suspicious (dysplasia), IV—carcinoma in situ, and V—invasive carcinoma.

● This case was adapted from a paper prepared for OTA by Anne-Marie Foltz, Yale University School of Medicine.

Some laboratories use as many as seven classifications and the names may differ. The first two classes are considered normal. Dysplasia and carcinoma in situ are considered "precancerous," while the last class is malignant.

For those with tests in the last three classes, the usual diagnostic procedure in the past and in areas without colposcopically trained physicians is a diagnostic-cone biopsy (removal of a section of the cervix). This is an in-hospital surgical procedure and carries a risk. Today, the usual procedure is a colposcopic evaluation (essentially, looking at the cervix with 15 x magnification) and biopsies (tissue samples) if there appear to be abnormalities on the cervix. Colposcopy, a relatively recent technique, requires some training.

Cervical cancer is widely believed to pass through three stages: dysplasia, carcinoma in situ and invasive carcinoma, with the process taking up to **35** years (62). This progression is supported by the evidence that the peak incidence of these conditions occurs at progressively higher ages, with the peak of invasive carcinoma at the ages of 60 to 64 (68).

There is also some evidence to the contrary. Invasive cancer has been found in women regularly and recently screened (Sandmire et al., 1976). The explanation may be that there are slow-growing tumors that pass through the three phases over 20 to 30 years, while others become invasive within a year. These findings are consistent with findings for lung and breast cancer (Charlson and Feinstein, 1974; Wells and Feinstein, 1977). Dysplasia has been found to regress, though probably not permanently (Stern and Neely, 1963); and in the few cases where carcinoma in situ has gone untreated, it has not necessarily progressed to invasive cancer (Spriggs, 1971).

This uncertainty about the natural history of the disease affects the efficacy of the test. It is difficult to evaluate efficacy if one cannot be certain what is being prevented.

The issue of the accuracy of the Pap smear test did not receive much attention when the test was disseminated. The accuracy of the test has been stated to be about 95 percent (ACS, 1975; Dickinson, 1972). However, this statement is misleading. In any condition with a low prevalence, such as cancer of the cervix, this statistic can hide a proportion of missed lesions (false negatives rate).

Recent studies have shown that cytologists read test results differently, particularly regarding carcinomas (Seybolt and Johnson, 1971; Lambourne and Lederer, 1973; Kern and Zivolich, 1977). Some of this variance occurs because the different classes are not clearly defined. Because of this variability, the number of lesions that are missed (false negatives) can vary from 2.4 to 40 percent (Husain, 1976; Coppelson and Brown, 1974). This variability also occurs among the pathologists who read the followup biopsies. Such misreading may lead to unnecessary hysterectomies, as this is the usual treatment for invasive carcinoma in situ in women who are not interested in future childbearing (Brudnell, 1973). However, hysterectomies also carry risks (see case 11).

Finally, false positive test results (those women with abnormal test results but no disease) are rarely reported in the literature, although these women may be subject to repeated test, biopsies, and perhaps hysterectomies with their concomitant personal and social costs.

The efficacy of the Pap test was not carefully studied before its wide diffusion. An efficacy experiment would have compared the rate of invasive cervical cancer and resulting death rates in a screened and treated population with those in an unscreened population. By the end of the 1950's, however, professional consensus endorsed the posi-

tive benefit of the screening procedure. A controlled clinical trial is difficult to carry out once a procedure is generally accepted as efficacious. Denying the supposed benefits to a segment of the population was believed to be unethical. However, it can be argued that if one can demonstrate substantial doubt of efficacy, it would be unethical not to study the technology.

In the absence of such a controlled study, other epidemiological methods have been used to estimate the value of Pap smear screening in cancer control. The two long-term screening projects on large populations have been those of Boyes in British Columbia and Christophersen in Louisville, Ky. The latter project has been supported by the National Cancer Institute (NCI). Reports from both studies, which have been operating more than 20 years, indicate that screening has led to a decline in mortality from cervical cancer. Christophersen has stated:

> That a decrease in death rates of the magnitude observed here is not to a major extent due to mass screening must be proved by a demonstration of a comparable decrease in an unscreened population. Such evidence has not been presented to date(68).

The decline in mortality was found to be significantly correlated with the intensity of screening in each State in the United States, but this may have been an artifact of intervening variables (Cramer, 1974). A more cautious analysis of screening and mortality, using Canadian data and controlling for socioeconomic variables, concluded that, at least for the age group 30 to 64, over the period 1960-62 to 1970-72, the intensity of screening had a significant effect on reduction of mortality (Miller et al., 1976).

It seems safe to say that screening seems, in some cases, to have had some effect on mortality. Proponents of an annual or frequent screening program cite the preventability of invasive cancer, the low cost of the test, the relation of screening rates to a fall in mortality, the need for frequent screening to catch fast-growing tumors, and the fact that any death from cervical cancer is preventable and therefore all women should be screened frequently (Guzik, 1977).

Opponents cite the low prevalence of the disease, the uncertainty of its natural history, and the accuracy of the test. Opponents may concede that screening has lowered mortality rates, but they point out that this seems to have occurred in areas such as Abderdeen, Scotland, where the screening intervals are not annual, but every 5 years (MacGregor, 1976).

In Canada in 1976, the Conference of Deputy Ministers of Health appointed a task force to evaluate the effects of screening. After **a** careful review of the scientific literature and in light of the costs of the program, they recommended in June 1976 that screening should be undertaken at the following intervals:

> A woman should have her first smear at age 18 if she is sexually active. If the initial smear is satisfactory, a second smear should be taken a year later.
>
> After that, further smears should be taken at approximately 3-year intervals until the age 35 and thereafter at 5-year intervals until age 60.
>
> Women at continuing high risk should be screened annually (62).

Because the Pap smear is a screening procedure, its efficacy and safety are not regulated by the Federal Government. The Center for Disease Control (CDC) and other Federal and State programs do regulate the quality of clinical laboratories that perform the cytological analyses, and Federal funds have been available to train cytologists. The Cancer Control Division of NCI in the past has supported research and screening pro-

grams through grants. Since 1974, in response to specifications in the 1974 Amendments of the Cancer Act, NCI's division of Cancer Control and Rehabilitation has supported the use of the Pap smear in 38 States through contracts with State health departments. These programs, which focus on reaching women who have never had a Pap smear, are being phased out. The Health Services Administration (HSA) of the Department of Health, Education, and Welfare (HEW) supports Maternal and Child Health Clinics and migrant health programs, which both offer the Pap smear. HEW also supports 4,500 family planning service sites that are required to provide Pap smears to all women using their services. The Pap smear is not covered as a benefit by the Medicare program, and its coverage by private insurance programs varies.

In summary, the Pap smear was widely diffused for 30 years without demonstration of its efficacy through controlled trial. Since then, its use has not been questioned, but its accuracy and the frequency of necessary screenings have been. Once the Pap smear was in widespread use, the very extent of use and professional consensus of its efficacy argued against carrying out a controlled trial. As the risks to women whose tests were found falsely positive by the Pap smear have never been seriously documented, it is possible that a controlled trial to examine that question may be of value. As case 1 illustrates, it is important that some method exist for bringing questions about the efficacy or safety of techniques technologies to the attention of investigators and public or private research policymakers.

Case 2: Amniocentesis*

Amniocentesis (from the Greek "amnion," the membrane surrounding the fetus within the uterus, and "centesis, " puncture) can be performed at various times during pregnancy for a variety of reasons. But it has come chiefly to refer to the most widely employed form of prenatal diagnosis. In this role, it is a method for obtaining a sample of the fluid that surrounds the fetus by inserting a hypodermic syringe through the abdominal wall into the uterus, generally at about 16 weeks gestation.

The procedure has been in existence for some time. Its use for discovering fetal sex was first reported in 1956 (I), but it did not come into wider use as a diagnostic technique until the early 1970's. The delay was partly for technical reasons: it was necessary to develop ways of examining constituents of the amniotic fluid that would reveal a disease or defect in the fetus. Development of amniocentesis also depended, however, on an important political change taking place about then: the loosening of legal restrictions on abortion, culminating in the 1973 Supreme Court decision (Roe v. Wade, 410 U.S. 113) that made abortion before 24 weeks gestation a matter to be decided between a pregnant woman and her physician, without State interference. That is because the goal of amniocentesis is information about the fetal state—information that will lead to the prevention of many kinds of birth defects by preventing the birth of those afflicted by them, via abortion.

An additional reason for the somewhat cautious early development of amniocentesis was concern about its safety, as it involved direct invasion of the uterus, and the risk to either mother or fetus was unknown. The National Institute of Child Health and Human Development (NICHHD) coordinated a study that pooled data on more than 1,000 cases of amniocentesis from nine major medical centers that were pioneering the technique; the results of that study were announced in the fall of 1975 and published a year later (2). The findings, since confirmed by other studies elsewhere in the world, were that amniocentesis was both safe and accurate. The difference between the rate of spontaneous

*This case is adapted from a paper prepared for OTA by Tabitha M. Powledge of the Hasting Center.

abortion among women who had undergone amniocentesis and the control group women who had not was not statistically significant, and maternal complications—vaginal bleeding, for instance—were minor. Followups of babies born after amniocentesis, at birth and at the age of 1, reveal no differences between them and other babies; longer term followups will, of course, have to await the passage of time. At this point, however, the technique (assuming it is done by qualified people) appears quite safe for both mother and baby—an important consideration, because in about 96 percent of amniocentesis cases, the tests will reveal no abnormality and the pregnancy will therefore be brought to term. The study also revealed that the error rate in diagnosis was substantially below 1 percent, Like safety, accuracy is an exceptionally important consideration in amniocentesis, as a wrong diagnosis will usually lead either to the birth of an unwanted, afflicted child, or the abortion of a wanted, unafflicted one.

A variety of tests can be performed on amniotic fluid, and others will probably be developed. Fetal cells obtained from the fluid can be laboratory-cultured and karyotypes (pictures of the fetal chromosomes) prepared from them; this procedure takes several weeks. The cells can also be examined for a variety of very rare biochemical abnormalities. Other constituents of amniotic fluid, such as hormones, can also give information about the fetal state. One expanding area of amniocentesis is the assessment of amniotic fluid α=fetoprotein, which is diagnostic of several kinds of birth defects, particularly the neural tube defects anencephaly and spina bifida.

Candidates for amniocentesis are drawn from groups of women thought to be at higher-than-average risk for bearing a child with a birth defect. This can sometimes be (and usually is, in the case of the rare biochemical abnormalities or sex-linked disorders) because she has previously borne an afflicted child. But the largest number of amniocenteses is performed on women over 35 (or, in some places, over 37 or 40), who are statistically at higher risk than younger women for bearing a child with a chromosome abnormality, particularly Down's syndrome, the most important single cause of severe mental retardation. Amniotic fluid α = fetoprotein assessments are becoming an increasingly important part of amniocentesis, particularly because of pilot programs such as the one currently going on in Nassau County, N. Y., where the assessments are used to confirm the less reliable diagnoses of neural tube defects obtained via assessment of α= fetoprotein levels in the blood of pregnant women (3).

The procedure appears safe (except, of course, for the affected fetus), but is it efficacious? Diagnostic accuracy alone satisfies only one of the possible standards of efficacy (see chapter 2). Though in almost all cases the results of the tap will be negative and therefore provide prospective parents with months of relief from anxiety, a small preliminary study has revealed a high rate of depression among mothers *and* fathers in cases where an abortion followed a positive diagnosis (6). The parents under study, however, did declare that, despite its psychological effects, they would certainly repeat the procedure rather than bear a defective infant. The situation will probably be worse for those parents who are opposed to, or ambivalent about, abortion.

Another problem is knowing which is the "defined population" in which the efficacy of amniocentesis will be judged. The mother alone? The fetus? The entire family, whose resources may be spared by prevention of the birth of an affected child? Society, whose resources are also at stake when care for the chronically ill or retarded is involved? This latter point is important because, while amniocentesis is usually justified as a needed service to individuals, a strong second line of argument has been that it can relieve some burdens on society. A proposal emanating from the Columbia School of Public Health for a gradual four-stage program that would eventually reach all pregnant women at-

tempted to demonstrate that even such a massive program would provide huge monetary savings over the cost of institutionalizing those with Down's syndrome (8). On the other hand, that same money might be more efficiently spent on improvements in prenatal nutrition or delivery procedures that might reduce the amount of mild mental retardation that is much more widespread in the population and may, on balance, constitute more of a burden to society than Down's syndrome.

Some other considerations that either bear on the question of what constitutes usefulness (broadly defined) or demonstrate that "efficacy" (narrowly defined) can be only a partial measure of the usefulness of medical technologies are:

. The cost of amniocentesis is not trivial. It currently ranges between $300 and $500, depending on the amount of laboratory work involved. Increasingly, that cost is being borne by third parties, either insurance companies or the State.

. Widespread use of amniocentesis will require a large and expensive personnel training program; most laboratories doing this work are already operating at capacity. The labs will also have to be monitored. The Federal Government seems the logical focus of both training and monitoring, but that, once again, means the cost will be borne by society rather than the individual.

• Amniocentesis provides a nearly foolproof way of finding out the sex of a fetus. Though not often employed in this way in the past (except for diagnoses of sex-linked diseases), its use for the purpose of picking the sex of children is likely to increase as facilities expand and more private physicians are trained in the technique. This in turn may require an array of public policy decisions, at least on the question of whether or not such use of amniocentesis should be subsidized.

In summary, amniocentesis appears to be safe and efficacious. Complex ethical and legal issues surrounding the use of this technology must be taken into account in an evaluation of its societal usefulness. Amniocentesis is peculiar because it depends in part for its effectiveness upon the wide availability of abortion. The fate of amniocentesis is therefore tied to the abortion debate in this country. This case demonstrates the problem of viewing efficacy and safety as the sole determinants of appropriate use.

Case 3: Chicken Pox Vaccine

A successful vaccine produces, without harm to the recipient, a degree of protection that approaches the immunity that follows a disabling attack of chicken pox itself. A vaccine is a preparation of bacterial or viral material that has been inactivated or weakened. This material can stimulate the body's immune system and prepare it to attack the agents of the corresponding disease should they invade the body.

By preventing disease, rather than treating them or their symptoms, vaccines have averted suffering and saved lives. Immunization programs have reduced financial as well as human costs. Vaccines such as those against smallpox, measles, and tetanus have prevented a variety of infectious diseases.

Chicken pox is an infectious disease caused by the Varicella-Zoster (VZ) virus. Chicken pox is a common, usually mild, childhood illness. The United States recorded 154,248 cases of chicken pox in 1975, an increase of almost 13,000 from the previous year (388). In 1974, only 106 deaths were reported from chicken pox.

Early in 1977, Japanese investigators reported the development of a live, attenuated (weakened) virus vaccine against VZ. Twenty-six children who had been exposed to

chicken pox were given the vaccine after exposure. None of them developed clinical chicken pox. A control group of 19 exposed children was left unvaccinated, and all developed typical chicken pox. Blood tests demonstrated development of immunity to chicken pox by those given the vaccine. On the basis of preliminary evidence, the Japanese vaccine can be considered to produce immunity to VZ virus and to prevent the development of chicken pox in those inoculated (12,13).

A vaccine against chicken pox, which appears to be a definite possibility, might prevent thousands of cases and more than 100 deaths per year. Most of those deaths, however, probably occur in individuals at high risk, such as those with leukemia. The danger of chicken pox in high-risk individuals could be reduced by a well-organized program of passive immunization with gamma globulin, which contains antibodies against chicken pox produced by other infected individuals (48). *

The risks of the chicken pox vaccine are unknown. Some normal children could react adversely. Because high-risk individuals would be likely to have variable reactions to the vaccine, it could be expected to cause a certain level of morbidity and mortality in this group (48).

The most worrisome risk, however, is the possible effect of the vaccine on the rates of zoster, another disease caused by the VZ virus. Zoster (shingles) is usually a disease of adulthood. It occurs in persons who have recovered from chicken pox many years after their recovery. The percentage of individuals who develop latent infection** is unknown, although the rate of zoster in high-risk populations, such as those on chemotherapy for cancer, has been reported to be as high as 50 percent (6). Why latent virus causes disease years later is not known, and the relationship between chicken pox and zoster is not well understood (134).

The vaccine could postpone infection from childhood, when it is a mild illness, to adulthood, when it may be quite severe. This could occur if immunity produced by immunization of infants and children waned in adulthood. Although reimmunization might prevent this problem, persuading adults that they need a vaccine against chicken pox might be difficult. In addition, the attenuated virus might itself become latent and cause infection years later. Because the latent period for zoster can be 10 to 30 years, results of vaccination would need to be studied for decades in order to establish the health benefit (48,388).

Furthermore, viruses related to the VZ virus have been shown to cause cancer in animals and have been related to some cancers in humans. This suggests some risk of producing cancer with such a vaccine (48).

Weighing these possible benefits and risks, Brunell has stated that “the mortality and morbidity produced by varicella (chicken pox) in normal children could hardly justify a major effort to eradicate varicella” (48). The National Institute of Allergy and Infectious Disease (NIAID) agrees regarding the use of live vaccine in normal children. Carefully controlled trials in children with cancer or leukemia under conditions of isolation may merit consideration. A killed or inactivated vaccine, such as a subunit vaccine free of nucleic acid, is also a possibility for varicella as for other herpes viruses (388).

‘*Passive* immunization refers to injection into the patient’s body of antibodies derived from another source, human or animal. *Active* immunization occurs when the antibodies are produced by the patient due to injection of a vaccine. *Passive* immunization of gamma globulin, therefore is not a full substitute for vaccination.

* *The state of continuing viral infection without clinical illness is referred to as “latency.”

NIAID has shown that a drug, adenine arabinoside, reduces the mortality of herpes encephalitis, thus offering promise that a drug efficacious against VZ virus may be developed. A natural antiviral substance, interferon, has been shown to limit the severity of VZ infections manifested as herpes zoster.

The live chicken pox vaccine is a case where efficacy can be predicted, but long-term risks in normal children are unpredictable without studies that would take decades. The benefits, although positive, are relatively small, while potential risks are large. This concern about potential risks has led NIAID to decide not to test the vaccine in normal children. A live chicken pox vaccine for general use seems to be a technology that will work, but that will not be developed in this country unless further research makes it possible to minimize risk.

Case 4: Mammography

Mammography is a special X-ray examination of the breast with a machine designed for that purpose. It is used both as a screening procedure on apparently healthy females and as a diagnostic procedure in clinical situations to detect breast cancer and to aid in diagnosis of the disease. The recent controversy and studies concerning mammography relate to its use in screening, not in clinical diagnosis. This case, therefore, examines the use of this procedure only for screening.

Breast cancer, the most common cancer among women in the United States, represents 27.2 percent of all cancers in women. It is diagnosed in about 90,000 women annually, and every year about 34,000 women die of breast cancer (8). It is the leading cause of death among women 40 to 44 years of age. Its mortality and incidence rates increase with age. Its incidence in the United States has increased since the mid-1940's (8,76).

Studies carried out before current treatments for breast cancer were available estimated mean survival from onset of symptoms at about 39 months, Only 18 percent of affected women survived 5 years without therapy (304). Today the overall 5-year survival rate is approximately 60 percent. (The two percentages cannot be directly compared, because breast cancer is being diagnosed earlier, and in the earlier study, survival was dated from onset of symptoms.)

Efforts to improve survival rates have emphasized early diagnosis and treatment. If the cancer can be found and surgically removed before it metastasizes (spreads) to other organs, the survival rate is good: a 5-year survival rate of more than 80 percent. * This has led to recommendations for periodic breast examination by a physician, for monthly self-examination of the breasts, and for periodic mammography.

Mammography was first used on patients in 1913, began to have more clinical use in the 1930's, was considerably improved by Egan in the 1950's, and was widely used starting in the 1960's. In the early 1960's, M. D. Anderson Hospital in Houston, with support from the National Institutes of Health (NIH), carried out a clinical trial of Egan's technique of mammography for the diagnosis of breast cancer, X-ray findings were correlated with pathological diagnoses of breast cancer on normal breasts in 1,580 patients. The technique was found to be reasonably accurate, with a false positive rate of 7 percent and a false negative rate of 6 percent. These findings encouraged use of mammography in screening for breast cancer (70).

● Information supplied by the National Center Institute.

In the mid-1960's, Shapiro, Strax, and Venet conducted a controlled clinical trial to see whether annual screening for 4 years with clinical examination and mammography affected mortality from breast cancer (313). More than 60,000 women were divided into a study group and a control group (313). The study (usually referred to as the HIP (Health Insurance Plan of Greater New York) Study) found that "repetitive screening with clinical examination and mammography leads to at least a short-term reduction in mortality from breast cancer. Over the 7-year period of observation for which data are available, there were 70 deaths due to breast cancer in the total study group as compared with 108 breast cancer deaths in the control group" (312). In a recently completed 9-year followup, Shapiro found 90 breast cancer deaths in the study group and 128 in the control (310).

Early findings from the Shapiro study, which remain valid after 9 years of followup, led NCI and ACS to support and promote, beginning in 1973, a Breast Cancer Diagnosis Demonstration Project (BCDDP). This program has involved some 270,000 women, ranging in age from 35 to 74, from 29 participating screening centers at a total FY 1977 cost of $9.5 million (386).

The screening program consists of instruction in breast self-examination, an initial clinical history and physical examination, mammogram (for certain age groups), and thermogram. All are repeated annually for 5 years, with a 5-year observation period after completion of screening. The estimated cost for the a complete individual examination is $35 per year. By 1976, about 1,800 cases of breast cancer had been found, at an approximate cost of $11,000 per case.

Recently, however, the safety of mammography has been questioned. Radiation dose from mammography can be as high as 6.5 rads per examination. Bailar stated that the risk to symptom-free women of getting cancer from high exposures might equal or exceed the benefit of finding a cancer early that could not be found by physical examination (16).

This potential risk led NCI and ACS to appoint three* expert committees in 1976 to assess the risks and benefits of screening, particularly with mammography and physical examination. Breslow chaired a group that reanalyzed the HIP data and affirmed the lack of benefit for women under the age of 50 and substantial efficacy for those over 50 (383). Although the Breslow Committee noted also that radiation dosage from mammography had decreased because of technological improvements since the HIP study, the committee was unable to ignore findings of the HIP Study because it was the only controlled study that examined the important parameter: overall reduction, from screening, in the number of deaths from breast cancer.

Another group, chaired by Upton, reviewed evidence concerning health hazards to those screened. It focused on whether radiation to the breast can cause cancer of the breast. The committee concluded that it can. It argued that even small doses of radiation to the breast are risky. The committee postulated a l-percent increase in risk with a dose of 1 rad to the breast. Assuming this dose, application of mammography to the entire population would thus add six avoidable cases of breast cancer per rad per 1 million women per year, after a 10-year latent period (383). However, Upton also noted that new equipment such as that used in the BCDDP delivers a radiation dose under 1 rad, perhaps as little as 100 mrad (0.1 rad).

● The third group studied the pathology of breast cancer and will not be considered here.

The three groups together made a series of recommendations (383):

1. That rigorous attempts be made to keep radiation dose under 1 rad per screening examination;
2. That mammography for routine screening of women under 50 years of age be discontinued; and
3. That NCI support a clinical trial of mammography to furnish more conclusive evidence of its usefulness.

On August 23, 1976, NCI and ACS, in a letter to directors and coordinators of demonstration projects for breast cancer detection, quoted findings from a preliminary report of the Breslow and Upton groups and concluded: "We cannot recommend the routine use of mammography in screening asymptomatic women ages 35 to 50 in the NCI/ACS BCDDP at this time. However, in the face of a very small presumed risk for any individual woman, we do not recommend withholding mammography from a woman age 35 to 50 years if she and the physician agree that it is in her best immediate interest" (117). This recommendation was intensified in May 1977, when BCDDPs were told that mammography was to be used on women with personal or family histories of breast cancer.

Because of the controversy surrounding mammography, NIH held a 3-day conference on breast cancer screening in September 1977. Sixteen leading scientists, epidemiologists, physicians, and lay persons reviewed technical information, ethical issues, and other information (on the development of the BCDDP project) and heard testimony from a variety of groups. They also considered the report of a special group set up by NCI to review the BCDDP data (384), The panel concluded that "the only sound scientific evidence which demonstrates favorable benefit in breast cancer screening is derived from the HIP Study. " Because of this and because of the radiation risk, the panel recommended continuation of the screening program in women age 50 and over, but recommended limitations for younger women. For women 40 through 49 enrolled in the BCDDP, the panel recommended mammography for women having personal histories of breast cancer or whose mothers or sisters have such histories. This was consistent with existing NCI guidelines initiated in September 1976. For women below 40, mammography was recommended only for those with personal histories of breast cancer (385). This recommendation has been incorporated into recent BCDDP guidelines.

The Bureau of Radiological Health of the Food and Drug Administration (FDA) has the responsibility to regulate X-ray machines used in mammography. Most observers agree that the radiation dose from use of mammography in the general community is too high. FDA and NCI, under an interagency agreement, are attempting to decrease the exposures used in State-operated, community-level programs of mammographic screening.

Third-party payment programs, including Medicare and Medicaid, cover mammography for diagnostic purposes. Screening is not consistently covered.

In summary, mammography is a screening tool for early detection of breast cancer that has been widely used as a result of studies in the 1960's. Questions about its safety have recently been raised, and it has become a controversial technology. Many believe that technological improvements make it efficacious and safe for all women, but there is no scientific information derived through controlled studies to support such a view. Most, including the NCI panel, believe it has been shown to be efficacious for women over the age of 50 and should be used routinely for that group. Because existing information did not adequately answer the question of net health benefit, NIH collected all available information and conducted the exercise described above. Not all controversy

33-496 0 - 78 - 4

was settled; therefore, future studies will probably be necessary, especially concerning the question of benefit in the 40 to 49 age group using the modern mammographic equipment. Mammography is an example of a medical technology that was being widely diffused before questions about its safety began to countervails conclusions about its efficacy, leading to a scientific controversy that may yet strike the proper balance for society.

Case 5: Prophylactic Oral Antibiotics in Elective Colon Surgery

Prophylactic use of antibiotics, the routine administration of such drugs prior to surgery to prevent postoperative infection, is very common in surgery on the abdomen. Each year approximately 217,000 persons undergo surgery on the intestines for such conditions as cancer of the colon (large intestine), polyps, and chronic ulcerative colitis (374).

A common complication of bowel surgery is contamination of the incision (wound) by bacteria normally found in the gut. Such contamination can lead to abscess formation, generalized sepsis (infection) with serious morbidity, and even death. Antibiotics began to be used to prevent postoperative infection shortly after their introduction in the late 1940's. Specific antibiotics have been found to destroy certain types of bacteria; thus, knowledge of the types of bacteria in the gut and the types of bacteria that cause wound infections permit identification of antibiotics that might be efficacious in preventing infection. Such antibiotics can be administered in several ways: intravenously (injected into the bloodstream) during, before, or shortly after surgery; orally a few days prior to surgery; applied locally following surgery; or combinations of these methods.

In the early 1960's, the Ultraviolet Light Commission found that patients who received prophylactic antibiotics had a higher incidence of wound infections than those who did not receive antibiotics (251). Since then, numerous studies of this technology's efficacy and safety have been conducted. Stone found, however, that "there are approximately 50 poorly founded and retrospectively reviewed 'testimonials' for every one controlled and statistically significant study" (326).

Clinical studies have serious problems themselves. Everett, for example, found no change in the incidence of wound infection when he used only neomycin (114). Because of the bacteria that neomycin inhibits, however, it could be expected to be only partially effective. Studies headed by Rosenberg (291) and Sellwood (307) with partially effective antibiotics found a significant decrease in the rate of infection, but the rate of infection in their controls was so high that their results must be viewed with caution. Barker used a combination of antibiotics now believed to be an ineffective dose (22). Nichol's study found no wound infections in patients given oral antibiotics, but his group was small and he did not use a double-blind experimental design (258).

The most rigorous study was Washington's, a prospective, randomized, double-blind study (409). A single surgeon performed the surgery in the study. The study found that a rational combination of oral antibiotics does reduce the rate of postoperative wound infections. Moreover, the treated group did not have serious postoperative complications because of the use of antibiotics.

The prophylactic use of antibiotics, in certain combinations and under controlled conditions, has thus been shown by one study to be efficacious and safe. The question of efficacy and safety, however, is rarely "settled for all time. " There still has been no widely accepted demonstration that systematic use of antibiotics prevents the complications

of elective colon surgery (397). Washington's study needs to be replicated and various combinations of antibiotics and methods of administration need testing.

FDA certifies antibiotics for both prophylaxis and for treatment. Thus, any approved antibiotic can be used prophylactically. To guide future decisions, the Veterans Administration (VA) is beginning a study to compare oral antibiotics with those given by injection. Use of antibiotics is covered under all Government medical care programs, including Medicare and Medicaid, and providers are reimbursed for using prophylactic antibiotics in bowel surgery. Reimbursement policies have not changed over the years, despite questions about such use.

This case study illustrates a technology whose use has been based not on testing but on surmise. After one study raised questions about the usefulness of such prophylactic antibiotics, however, a number of clinical trials were carried out, some with Federal support. Most of these trials have been inconclusive, because of methodological problems. One recent study allows a tentative conclusion that prophylactic antibiotics are useful in colon surgery. However, the many variables involved in the situation lessen the impact of any single study and also make complete assessment very difficult.

Case 6: Skull X-Ray

X-ray of the skull is a standard diagnostic procedure widely used in the United States for a variety of conditions. Approximately 17 million skull films were taken in this country in 1970, in the course of about 4.2 million skull examinations (362) (each skull examination includes multiple skull X-rays). In 1977, an estimated 5.7 million skull examinations were carried out. * The major corrective treatment for abnormal conditions within the skull is surgery; about 70,000 intracranial operations are done per year. * * The validity and reliability of skull X-rays have been studied extensively, but, according to Weinstein, Alfidi, and Duchesneau, their use produces an "extremely low yield of meaningful information that will contribute to the potential diagnosis or alter the course of therapy" (414).

Skull X-ray is used widely (in conjunction with physical examination, history, etc.) as a screening tool, especially in case trauma to the head, to determine if injury has taken place. An estimated 20 to 30 percent (0.8 million to 1.3 million skull examinations) of these examinations each year are done to evaluate head injury (28). Bell and Loop studied the use of skull X-rays for trauma by two hospitals in 1969 and 1970. They reported that 93 fractures were found in 1,500 skull examinations, or 1 in every 16 skull series. They found that the physician's evaluation of the patient was relatively accurate, especially with more severe injuries. Furthermore, only 28 of the 93 patients with skull fracture (30 percent) had therapy altered because of the demonstrated fracture; in those cases, skull fracture led either to prophylactic antibiotics or to surgery. Bell and Loop stated that "unsuspected fractures may be associated with less trauma and less disability, and perhaps seldom need to be demonstrated. " They also found that 20 percent of examinations were done for "trivial injury" and that another 34 percent were done to protect against possible malpractice suits (28).

Other findings also suggest excessive use of skull X-rays. Lusted and his coworkers found that about 16 percent of skull X-rays were ordered even when the physician reported certainty about the diagnosis (220). Jergens, Morgan, and McElroy studied a large emergency room and found the same situation as reported by Bell and Loop. Less

*Information furnished by the Bureau of Radiological Health, HEW.

• ● Information furnished by the National Center for Health Statistics, HEW.

than 1 percent of skull X-rays were positive and 19 percent were ordered for medico-legal reasons. They also noted that many examinations were done at the request or the demand of the patient.

Skull X-rays have little direct impact on therapy because the underlying brain damage, not fracture, is the critical variable for treatment—and brain damage does not appear in X-rays. Roberts and Shopfner state, "physicians can instruct patients and lawyers that head trauma causes injury to all cranial structures, including the brain, blood vessels, bone, and scalp, but that bone fracture almost never has any bearing on the patient's need for treatment and hospitalization" (288).

The apparently limited benefit from skull X-ray also needs to be weighed against the risk of exposure of a large population to radiation. One skull X-ray causes about 330 milliroentgens exposure; with an average of four X-ray exposures per skull examination or series, the average exposure to the individual is about 1.3 roentgens (362). Although no specific risk can be assigned this amount of radiation (2), risks of radiation should be minimized whenever possible.

The costs of skull X-rays also need to be considered. In 1970, when a skull series cost about $30, the aggregate cost was about $120 million (28). By 1977 the cost for a skull series has risen only to $39, yet the aggregate cost was $221 million. * Weinstein, Alfide, and Duchesneau comment: "We do not wish to imply that all skull roentgenograms are contraindicated. However, millions of dollars could be saved annually if skull roentgenograms were obtained only when indicated" (414). Bell and Loop developed a list of indications for skull X-ray in trauma and found that 29 percent of all those given skull X-rays did not meet any of their criteria (28).

The Federal Government is not supporting any clinical trials on skull X-rays. However, the FDA's Bureau of Radiological Health has supported the development of criteria for appropriate use of skull X-ray. Phillips (274) developed such criteria, based on the work of Bell and Loop (28), in 1973. Beginning in 1975, the high-yield criteria were applied in the emergency room of the University of Washington Hospital to all cases of head trauma. Despite a compliance rate of only 55 percent by physicians, the number of skull examinations for trauma decreased 39 percent from the previous year.

This result was encouraging enough that the Bureau of Radiological Health has supported an extension of use of the criteria to 5,000 patients in Washington State, working through the Washington State Professional Standards Review Organization (PSRO) program. If the project is successful, it might be extended to all PSRO programs. The National PSRO office is following the experiment with great interest.

The X-ray machines used for skull X-rays are regulated by the Bureau of Radiological Health to minimize population exposure to ionizing radiation. Skull X-rays as ordered by a physician are provided under all Federal programs for medical care and reimbursement. PSROs do not generally review skull X-rays.

In summary, skull X-ray is a technology with recognizable risks and a large financial cost. Whether the technology can be regarded as efficacious depends on the level of diagnostic efficacy at which it is being evaluated (see chapter 2). For example, is it efficacious in terms of accurate diagnoses? Its effect on diagnosis and patient outcome appears to be limited; thus, it is of low efficacy by those criteria. This case, therefore, points out the importance of specifying which level of diagnostic efficacy is being used in evaluating the usefulness of a diagnostic technology. Careful studies of indications for use

*Information furnished by the Bureau of Radiological Health, HEW.

could improve the application of the technology. At present, skull X-ray appears to be overused. If this is the case, then aggressive policies to decrease such use, especially in trauma cases, could decrease wasted expenditures and prevent unnecessary radiation exposure.

Case 7: Electronic Fetal Monitoring

Fetal monitoring is the continuous observation and recording of biological variables considered to be reliable indicators of a fetus' condition. In practice, fetal monitoring is done during labor, and has traditionally involved monitoring of the fetal heart rate by a nurse using a stethoscope (auscultation). An electronic device for fetal monitoring is a recent innovation, and its use has been growing. In "indirect" (267) or "noninvasive" (336) monitoring, the fetal heart rate and uterine contractions are monitored by sensors placed on the woman's abdomen. In "direct" or "invasive" monitoring, an EKG (electrocardiogram) electrode is attached to the head of the fetus through the vagina (336). In direct monitoring, a small needle is often inserted into the fetal scalp to sample fetal blood (336). In addition a catheter is usually passed into the uterus to obtain information about the frequency, duration, and intensity of uterine contractions (63).

The rationale behind fetal monitoring, in general, and electronic monitoring, specifically, is that the condition of the fetus can deteriorate rapidly during labor (56). So-called "fetal distress" can lead to mental retardation or even death. About 7,500 infants annually die during labor in the United States (63). Another 44,000 individuals are born mentally retarded each year (281). If fetal distress is discovered by changes in the fetal heart rate or in the acid-base balance of fetal blood, Cesarean section might save the life of the fetus or prevent brain damage.

Obstetricians and neonatologists (those specializing in medicine concerned with the newborn) believe that electronic fetal monitoring (EFM) is markedly superior to monitoring done with a stethoscope (56,169,182,281,327). Some propose its use in all deliveries. Several experts have suggested that monitoring by stethoscope is essentially useless (63). A report from the Pan American Health Organization states that "the appraisal of fetal condition by cardiac auscultation and palpation of the uterus is a less accurate, not continuous, time-consuming, and fatiguing method. In the majority of cases it does not enable the early detection of fetal distress" (56).

The belief of obstetricians in the efficacy of EFM is largely based on falling newborn mortality rates in institutions where EFM has been introduced (27,110,168,173). A typical experience is that reported by Quilligan and Paul, who found that the neonatal death rate at their institution fell after introduction of electronic monitoring (281). However, other changes were occurring in obstetrical practice besides EFM during the same period (209,269), and important changes were taking place in the general health of pregnant women. Better nutrition has been provided to pregnant women, and widespread contraception and abortion have changed the age at, and conditions under, which many give birth, leading to a better outcome (145,262,338). Wennberg also analyzed this question by examining hospitals in Vermont (420). He found a 30-percent decline in neonatal mortality rates from 1969 to 1974 at university hospitals where EFM had come into use—and a similar decrease in death rates in other hospitals in Vermont without any changes in obstetrical practices.

A few reports from institutions have analyzed their results by birthweight. Several investigators have found that low-birthweight newborns account for more than half of neonatal mortality (168,269). When results are analyzed by birthweight, much of the

change in perinatal mortality is in this group. Beard (1975) found a striking decline in neonatal deaths in premature infants with electronic monitoring. Wennberg also examined the factor of birthweights in Vermont, and found that perinatal death rates have fallen markedly in newborns under 2,500 grams over the past decade, while the death rates in normal-sized newborns have remained unchanged. The implication is that modern obstetrics has been of value to the small-birthweight infant, but of little benefit to the normal-sized infant.

The question of efficacy of EFM could be studied by controlled clinical trials, which would report on the results of long-term followup of children born with and without fetal monitoring. The three clinical trials that have been done so far have looked only at short-term outcomes, all in high-risk women. Two trials carried out in Denver found no benefit when EFM was compared to nurse monitoring (159,160). Efficacy of monitoring was measured by infant outcome on a variety of measures, including neonatal death and neonatal nursery morbidity. In fact, outcomes were the same in the two groups, but it was observed that EFM intruded on the process of birth and that it depersonalized care. The flashing lights of the monitors adjacent to the bed and the sound of each fetal heartbeat disturbed the mothers. Haverkamp also noted that "very close physical contact with the patient was necessary for the nurse to auscultate fetal heart tones adequately. This was not true to the same degree with the monitored group. Nursing attention to the gravida (pregnant woman) with respect to maternal comfort, emotional support, and 'laying on of hands' could have a significant impact on the fetus. " However, the trial excluded low-birthweight babies. A similar trial, from Australia, found substantial benefit but included low-birthweight infants (283). All three trials had methodological problems, particularly, the failure to use a research design that would minimize the influence of investigator bias on the results. The findings of the three trials are also consistent with a small degree of benefit.

However, Neutra and his coworkers used the. data from several years' experience at a large hospital to develop a statistical model of monitoring, and did find a modest benefit. In their model, monitoring 27 percent of labors with demonstrable risk factors would avert 80 percent of the potentially preventable neonatal deaths. Thus, clinical trials have not demonstrated clinical benefit, but clinical experience does suggest benefit for low-birthweight infants. It has also been claimed that monitoring prevents fetal brain damage (281), but there is no evidence of such benefit (145).

Electronic monitoring has its risks. Scalp abscesses and lacerations of the fetal scalp and perforations of the uterus can occur (63,238,267). Uterine infection can occur from the catheter (135,208). Also, practices associated with use of fetal monitors may induce the very fetal distress they are meant to detect. * Before an internal monitor is inserted, the amnionic sac must be ruptured, which may cause abnormally strong contractions that increase fetal stress. Effective use of the external monitor, on the other hand, requires that the woman remain still, which may have the effect of prolonging labor (studies show that frequent changes in position, and upright positions, speed labor). Furthermore, if a woman lies on her back to avoid disrupting an external monitor, the weight of the fetus in the uterus may constrict circulation in the aorta and vena cava and cause depression of the fetus, and maternal blood pressure, or both ("vena cava syndrome").

The most important risk to both mother and child, however, is Cesarean section and its risks. The Cesarean section rate has risen in the United States from 5.5 percent of deliveries in 1965 to 12.5 percent in 1976. There seems little question that this rise is

*Information furnished by the Food and Drug Administration.

associated with electronic monitoring. Many institutions report higher Cesarean section rates in monitored than unmonitored patients or increased Cesarean sections after introduction of EFM (69,130,131,269,314,327,420). In the first Denver controlled clinical trial (160), the Cesarean section rate was 6.6 percent in the nurse-monitored group and 16.5 in the EFM-monitored group. The rise seems to be associated with an increase in the diagnosis of fetal distress that follows monitoring, although other changes in obstetrical practice also contribute somewhat to the rise (11,185,344).

The use of monitoring has been increasing rapidly. By the end of 1972, an estimated 1,000 fetal monitoring systems were in use in the United States (267). It is probable that all obstetrical services soon will have monitoring capability and that more than half of the approximately 3 million deliveries a year could be monitored electronically.

Sales of monitoring equipment reached $25 million in 1976 and may reach $40 million (in today's dollars) by 1986 (219). Estimates of the added cost per delivery of EFM range from $35 to $50 (281) to $75 (157). Thus, if electronic monitoring were used in every delivery, it could cost society $200 million or more.

Delivery by Cesarean section increases the cost of delivery from $700 to $3,000 (157). Thus, if half of the increased number of Cesareans are attributable to normal fetal stress that is interpreted as fetal distress, $175 million has been added to the national health bill from Cesarean section associated with use of electronic fetal monitoring. This estimate does not include the cost of death and morbidity of mother and child from monitoring and Cesarean section.

Legal issues complicate the use of electronic fetal monitoring. The risks raise the possibility of malpractice suits. On the other hand, with strong professional support for electronic monitoring, physicians who do not use it may also face malpractice suits.

NICHHD of NIH is funding one study of electronic fetal monitoring. The Office of Maternal and Child Health of HSA is supporting a study by Haverkamp and his coworkers comparing nurse monitoring, electronic monitoring, and electronic monitoring with fetal scalp sampling.

Electronic fetal monitors are regulated by FDA under the Medical Device Amendments of 1976. Several local and State governments have taken steps toward requiring that all hospitals with maternity care units provide electronic and biochemical fetal monitoring as well as trained personnel to carry out the monitoring. The Health Department of New York City has already made such a recommendation (29).

Electronic monitoring is usually covered under third-party reimbursement programs, including Medicaid. Other Federal programs also provide it. For example, the Office of Maternal and Child Health of HSA distributes formula grants to States to support maternal and child health clinics whose intensive care units for women and infants considered to be in "high-risk" categories provide electronic fetal monitoring.

In summary, although many believe that electronic fetal monitoring is useful, its relative efficacy and benefit have not been established. Two controlled studies indicate that monitoring by nurses may be equally efficacious and provide additional benefits; a third finds EFM to be of some relative benefit. Moreover, fetal monitoring may be associated with considerable risks and financial costs. It is a technology that may well have been diffused prematurely. It is an example of a technology for which guidelines on appropriate indications for use might be needed. Guidelines could suggest what types of patients and delivery situations would result in benefits exceeding the possible risks.

Case 8: Surgery for Coronary Artery Disease

Coronary artery disease is caused by narrowing and blocking of the arteries that supply blood to the heart. The blockage results from arteriosclerosis (hardening of the arteries). The most common manifestations of coronary artery disease are myocardial infarction (heart attack or coronary), angina pectoris (severe temporary chest pain), and sudden death.

Coronary heart disease is the number one cause of death in the United States. In 1975, it was responsible for 642,719 deaths. The same year an estimated 4,120,000 Americans reported a history of heart attack and/or angina pectoris. Arteriosclerotic heart disease was the most frequent condition diagnosed for patients at the time of discharge from hospitals in this country in 1968 (210).

For more than half a century, surgeons have believed that an efficacious surgical approach to coronary artery disease is possible. Prior to the modern bypass operation, five different operations were developed and advocated enthusiastically (279). Although all five operations were ultimately abandoned as of no value, initially they were alleged to be efficacious, with reports in the medical literature claiming "objective" evidence of benefit. These operations were accepted and diffused by many members of the medical profession on the basis of experiential evidence. Other physicians usually preferred careful medical management and sound advice on how to conduct one's life, with surgery as a second line of defense.

For example, in the 1950's a surgical operation called internal mammary artery ligation was widely advocated by a small number of surgeons for improving blood supply to the heart. In retrospect, this procedure has little scientific rationale. The mammary artery is tied surgically. Because this artery is near the heart, surgeons hoped that this action would force blood to flow through other arteries in the vicinity, including coronary arteries.

In 1958 and 1959, two randomized, controlled clinical trials were conducted by, respectively, Cobb and Diamond (51). Patients were assigned randomly to control or operative groups, and the control group was given a sham operation, * in which the internal mammary artery was surgically exposed, but was not ligated. Both groups of patients reported relief from anginal pain and increased tolerance of exercise. As a result of these trials, the operation was largely abandoned. That both groups benefited suggests a strong placebo effect in the treatment of angina.

The experience with prior surgical operations for coronary artery disease points out that: (1) initial enthusiasm for, or belief in, an operation, based on current medical concepts, did not assure or predict results; (2) experiential evidence (anecdotal) led physicians to the false conclusions that the operations were successful; (3) with the exception of the internal mammary artery ligation operation, no truly objective (scientific) assessments of efficacy were made; (4) the operations were diffused without prior testing of efficacy or evaluation of safety; and (5) physicians reported dramatic relief of symptoms (angina) for all operations, demonstrating that a double-blind study is often necessary for evaluation of symptomatic response to technological intervention.

Coronary bypass surgery was introduced in the early 1970's. In this procedure, a graft is put on the coronary artery to bypass the constricted portion of the artery. This procedure has become the primary surgical approach to treatment of coronary artery dis-

*Sham surgery in a clinical trial would most likely not be possible today because of ethical considerations.

ease (51). Approximately 25,000 operations were performed in 1973 and at least 70,000 in 1977. Yet the benefits of coronary bypass surgery have not been clearly demonstrated. Claims that the operation prevents death remain largely unproven (73). Nonetheless, one proponent was quoted as saying that the United States should prepare to do 80,000 coronary arteriograms a day to screen for coronary disease. Coronary arteriogram is a special X-ray examination of the coronary arteries that is used to gather information useful in deciding whether to perform the bypass surgery. Such a widespread diagnostic program would itself cost more than $10 billion (162).

Coronary bypass surgery seems to give excellent symptomatic relief from angina pectoris. It is reported that 70 percent of patients evaluated 1 to 60 months after surgery are initially completely relieved of angina (210), but the improvement diminishes with time. However, the placebo effect mentioned above needs to be kept in mind because: 1) the initial results are similar to previous operations; 2) nonsurgical treatment also produces good results; and 3) the methods of evaluation of symptomatic relief are experiential.

Several important clinical trials of coronary bypass surgery have been conducted. From 1970 to 1974, VA conducted a randomized prospective cooperative trial that compared the efficacy of medical to surgical therapy for patients with stable angina pectoris (398). Of the 1,015 patients in this study, 113 were found to have a significant narrowing to the left main coronary artery. On followup of this group, those treated by surgery had a better survival rate. In the main study group, however, there was no difference in survival between medically and surgically treated patients. Surgery appears to have little effect on mortality except in a small group of patients.

The National Heart, Lung, and Blood Institute (NHLBI) is sponsoring two trials of coronary artery surgery. One compares medical to surgical therapy for patients with unstable angina. To date, the mortality rate is low and comparable for both groups, but the surgically treated group has had an incidence rate of myocardial infarction higher than that of the medically treated group. The second trial will resemble the VA study. Its results are not yet available. Three other randomized controlled trials in this country show no difference between surgical and nonsurgical groups (197,235,306).

Many advocates, convinced of the efficacy of the surgery, have declined to participate in clinical trials. The same advocates argue that the results of clinical trials may not be valid because some of the most skillful surgeons have declined to participate in the trials (200).

The risks of coronary bypass surgery are similar to those of any major surgery. The hospital mortality rate for patients undergoing such surgery is reported between 0.3 and 8 percent, with a usual range of 1 to 4 percent. However, only good results are published, generally, and the operative mortality rate derived from a large number of hospitals providing comparable data was 4 percent in 1976. * Other complications include myocardial infarction during surgery, in about 7 percent of patients (210).

The total cost of a coronary bypass procedure averages $15,000, so that aggregate costs in 1977 were more than $1 billion. Most of this amount was paid by third parties. Medicare and Medicaid programs reimburse for such surgery when considered by a physician to be medically necessary, On a per capita basis, Health Maintenance Organizations (HMOs) use the operation at less than one-half the national rate, and in Western Europe the rate is about 7 percent of the rate in the United States.

*Source: Commission on Professional and Hospital Activities, Ann Arbor, Mich.

Coronary artery bypass surgery is based on a scientific rationale and may be of measurable benefit to some patients. It is usually performed for angina pectoris and appears to give substantial relief from symptoms, but the extent to which this relief is an effect of surgery is not known. Limited studies suggest that coronary bypass surgery improves life expectancy significantly for only a small number of patients, with a particular type of coronary artery disease. Controlled studies have shown no improvement in life expectancy for patients studied.

Case 9: Tonsillectomy

Tonsillectomy is surgical removal of the tonsils, small bodies of lymphoid tissue in the throat. Tonsillectomy is the third most common operative procedure performed in hospitals in the United States. Approximately 884,000 tonsillectomies were performed in 1973 (374), and about 680,000 in 1976. Removal of the tonsils is by far the most frequent surgical procedure performed in hospitals for patients under the age of 15.

Tonsillectomy has been done throughout recorded history, with attempts at removal dating at least as far back **as** 600 B.C. (265). Before **antibiotics, it was** probably medicine's only weapon against serious complications of throat infections (tonsillitis). After 1900, refinement of surgical technique encouraged its wide application. The popularity of tonsillectomy peaked in the 1930's, and its use has gradually declined since then.

Despite its long history, tonsillectomy has not been well evaluated for efficacy. The inadequate design of published studies makes credible conclusions about its relative benefits impossible (37). Paradise has summarized problems of experimental design (264):

1. The selection of patients for surgery was not random.
2. Severity of tonsillitis varies within and between operated and control groups.
3. Indications for surgery were not stringent, so that many children with mild or no disease were subjected to operation.
4. Because of ethical considerations, children who appeared to the investigators most in need of surgery were excluded from studies and given the operation.
5. Postoperative evaluation was based not on direct examination of the children but only on information obtained from parents.

In part because of the lack of experimental knowledge, the attitudes of pediatricians and surgeons toward tonsillectomy vary greatly (264,328). Some believe it to be a useless procedure and routinely refuse to perform or recommend it. Others, impressed by cases of children whom tonsillectomy appears to have helped dramatically, continue to recommend it. Paradise, et al., have stated, "Differences among authorities aside, a history of recurrent throat infection remains the indication for tonsillectomy most commonly advanced by parents and invoked by physicians, and constitutes a principal criterion in current quality-of-care standards for the reasonableness of tonsillectomy" (266). Tonsillectomy is uniquely indicated when the tonsils are large enough to obstruct breathing or swallowing. Even accepting these indications for tonsillectomy, a significant number of physicians believe that many unnecessary tonsillectomies are performed (264,328).

It has been estimated that 30 to 40 deaths a year result from tonsillectomy (434). Other estimates run as high as 300 deaths per year. * Postoperative hemorrhage, either

*Personal communication, J. Paradise, M.D.

immediate or delayed, can contribute to the morbidity attributable to tonsillectomy. Psychological risks, although difficult to document, certainly exist. Some speculate that serious problems such as Hodgkin's disease can result years after tonsillectomy (265), but no long-term ill effects have been demonstrated convincingly.

The rates for tonsillectomy vary considerably. For example, one study found that rates of tonsillectomy varied from 20 per thousand to 5.6 per thousand depending on area of the country (348). Tonsillectomy is covered by most if not all third-party insurance plans, including Medicaid. A study of 22 States, encompassing more than 6 million Medicaid eligibles, showed markedly different rates of tonsillectomy by area of the country, varying from a high of 1,709 per 100,000 people in Nevada and 1,324 per 100,000 in Maine, to a low of 179 per 100,000 in Arkansas (348). The total cost of tonsillectomy in the United States is estimated at up to $500 million per year (434).

NIH funded a controlled clinical trial of tonsillectomy and adenoidectomy at the Children's Hospital of Pittsburgh in 1973. The Pittsburgh group has made a concerted effort to define carefully the group that would be admitted to surgery and to ensure that this group did in fact have repeated episodes of tonsillitis. A preliminary report from the study shows the importance of doing so, since most patients with histories of recurrent infections that were not well documented proved to develop relatively few episodes when followed closely (266). For patients actually admitted to the randomized clinical trial, careful followup of both the operated and control groups is being done. In March 1978, NIH funded the study for 3 more years.

In 1974, NIH sponsored a Workshop on Tonsillectomy and Adenoidectomy. Its participants concluded that a nationwide, collaborative, controlled clinical trial of tonsillectomy was indicated, modeled after the Pittsburgh study. More recently, NIH funded a group to assess the feasibility of such a multicenter trial. The findings of that group were presented at the July 1978 meeting of the Ad Hoc Advisory Panel on Tonsillectomy and Adenoidectomy. The panel did not reach unanimous agreement with the group's recommendation to go ahead with the multicenter trial.

In summary, tonsillectomy is a surgical procedure that has long held a place in medical practice, but its efficacy and indications for use are inadequately understood. Reliable and valid data are not available, and the practicing community has reached no consensus on its value. Available evidence seems to indicate that many unjustified tonsillectomies are performed, especially in some areas of the country. The major well-controlled study currently in progress in Pittsburgh may provide better data on the efficacy of tonsillectomy and its indications. However, developing better information is only the first step. After that, the cooperation of the practicing medical community will be necessary to bring medical care more in line with the new information.

Case 10: Appendectomy*

Appendectomy is surgical removal of the appendix, a small tubular extension of the intestine ordinarily located in the lower extension of the intestine in the lower portion of the abdomen. It is usually performed as treatment for appendicitis, inflammation of the appendix. Without treatment, some inflamed appendices perforate and release bacteria into the abdominal cavity. Such perforation can cause peritonitis, a generalized infection of the abdominal cavity that can threaten life.

*This case is adapted from material prepared for OTA by Richard Watkins, M. D., a member of the advisory panel for the study.

In 1973, approximately 350,000 appendectomies were performed in the United States (374), and 1,060 deaths from appendicitis were reported (436). Although physicians and the public believe in the efficacy of appendectomy (35,52), no controlled clinical trials have been carried out. A study in China of 955 cases of appendicitis treated without surgery reported two deaths (417,1). One trial of nonsurgical treatment in the Western World reported 471 cases and one death (75). Although one cannot generalize from these trials because of their small size and other factors, the reported appendicitis death rates from the trials are lower than the 1973 U.S. death rate for appendicitis (436). The number of deaths attributable to appendectomy itself is not known. If the risk of death is estimated to be between 0.01 and 0.1 percent, deaths from appendectomy in the United States would be between 35 and 350 per year.

Examination of the mortality rate from appendicitis over time raises questions about the effectiveness of appendectomy. Appendectomy was widely adopted after 1900. Appendectomies were performed at rates of about 400 appendectomies per 100,000 population in 1920, about 600 in 1930, and 800 in 1938 (80). The reported appendicitis death rate rose from about 10 deaths per 100,000 population in 1900 to 13 in 1920 and 15 in the early 1930's (80,211,430). Increasing mortality over the early decades of appendectomy has also been noted for Australia (120) and the United Kingdom (44).

In the 1930's and 1940's other therapies for appendicitis came into use, notably intravenous fluids, relief of abdominal distension by a tube passed into the stomach, and antibiotics. Several writers have attributed the subsequent fall in rates of mortality to those innovations (120,337). Mortality began to decline from its high of 15 per 100,000 in the mid-1930's to 10 deaths per ,100,000 in 1940 (75), two deaths per 100,000 in 1950 (389), and one death per 100,000 in 1960 (389). The appendectomy rate also fell from about 700 per 100,000 in 1940 (80) to 200 per 100,000 in 1965 (374).

The beneficial effects of antibiotics and other technologies might have obscured any effect of surgery on mortality in the 1940's and 1950's. Assuming that appendectomy generally prevents death, rates of death and rates of appendectomy should be inversely correlated (255). Both rates, however, have continued to drop, the mortality rate falling to 0.9 deaths per 100,000 in 1965 (390) and 0.5 deaths per 100,000 in 1973 (374), and the appendectomy rate falling to 160 appendectomies per 100,000 population in 1973 (374,436).

The rates of appendectomy for regional U.S. populations for 1965-73 vary from 100 to 620 per 100,000 (100,214,421,422). Rates among Federal employees using different health care systems contrast sharply. In 1968, Federal employees who received medical care from 14 prepaid group practice plans underwent appendectomy at the rate of 110 per 100,000, while Federal employees enrolled in Blue Shield underwent appendectomy at the rate of 210 per 100,000 (272).

The Group Health Cooperative of Puget Sound, a large prepaid group practice with an age and sex composition similar to that of the United States as a whole, had an appendectomy rate of 105 per 100,000 population from 1970 to 1976, and an appendicitis mortality for the same period of 0.24 deaths per 100,000 population. These rates may be compared to an appendectomy rate of 160 per 100,000 and 0.5 deaths from appendicitis per 100,000 for the United States as a whole in 1973 (374,436). Group Health Cooperative physicians tend to observe the patient when the diagnosis of appendicitis is dubious (411). Possible, mild inflammation of the appendix subsides during observation, and surgery is avoided. Recently a group of surgeons at Johns Hopkins University found that observation in dubious cases reduced their overall appendectomy rate by almost one-third without an increase in perforation (423). The use of more discriminating criteria for appendectomy appears likely.

The cost of appendectomies in the United States is estimated at more than $350 million annually (28). Much of this cost is covered by third-party payers, both public and private. Appendectomy is a standard benefit of almost all health insurance programs, including Medicare and Medicaid.

Thus, appendectomy is a costly technology with the standard risks associated with surgery. The relative benefits and risks of treating appendicitis through surgery or other treatment have not been fully evaluated. For example, there is strong evidence suggesting that appendicitis may be treated with substantially fewer appendectomies without increased loss of life. Thus, a controlled clinical trial of the nonsurgical or delayed-surgical approach to treatment of certain categories of patients with evidence of appendicitis might be warranted.

Case 11: Hysterectomy

Hysterectomy is surgical removal of the uterus. It can be performed by either gynecological or general surgeons; indeed, legally, by any physician. The National Center for Health Statistics (NCHS) estimates that 678,000 hysterectomies were performed in the United States in 1976. At a rate of 622.2 hysterectomies per 100,000 females per year, this major operation is performed at a higher rate than any other. If such a rate continued into the future, more than half of U.S. females would have had their uteruses removed by age 65 (49). Moreover, the rate increased approximately 25 percent from 1965 to 1976 (348). In the late 1960's the hysterectomy rate in the United States was more than twice as high that of England and Wales (50).

These facts helped lead to allegations that hysterectomies are carried out unnecessarily in many patients. However, there is no clear-cut definition of what is necessary; nor are the indications known for those hysterectomies that were performed.

Hysterectomy is performed for a variety of conditions, including premalignant states and localized cancers (see case 1), descent or prolapse of the uterus, and obstetric catastrophes, including bleeding and septic abortion. Recently, indications for the operation seem to have been broadened beyond those traditionally accepted. Functional problems and conception control have become common indications. Cole argues that the differences in national rates and the increase in the rate of hysterectomy in the United States are a result of "prophylaxis, " that is, to prevent later cancer or pregnancy. The reasoning is "based on the rationale that if a woman is 30 or 40 years old and has an organ that is disease-prone and of little or no further use, it might as well be removed" (77).

Hysterectomy has risks. Cole and Berlin estimate a mortality rate of 0.06 percent, or 600 deaths per 1 million women operated on (78). Operative morbidity, although difficult to quantify, also exists. About 30 percent of women have postoperative fever and 15 percent require transfusions, which introduce some risk of hepatitis. Other potentially important health losses are less obvious. Hysterectomy appears to affect ovarian function, even when the ovaries are left intact. It has been postulated that if estrogen (female hormone) levels are affected by hysterectomy, higher rates of coronary artery disease could result (78). Even a l-percent increase in death rates from coronary disease would offset any possible gain from preventing cancer (77). The psychological response to hysterectomy may be another major problem. Several studies have found psychiatric disturbance, including severe depression, in women after hysterectomy. Despite methodological problems, these studies seem to indicate a significant amount of disturbance. Notman believes it may be difficult for a woman to adjust to the loss of reproductive poten-

tial, but emphasizes the need for well-controlled studies of the emotional consequences of hysterectomy (259).

Cole has analyzed the benefits that could be derived from carrying out hysterectomies on 1 million women at age 35. Assuming a conservative 600 deaths from the operations, the million women would overall have a slightly longer life expectancy as a result of surgery. Only the 1.3 percent of women who would have died from cancer of the cervix and uterus would benefit, with an average of 14.3 years of life each (77). These calculations assume a constant rate of occurrences of cancer of the cervix and uterus.

In economic terms, Cole estimated that 1 million hysterectomies would cost $2.9 billion, and would result in savings of $1.4 billion, including 35,000 cases of cancer. He concludes on the basis of his analysis that the benefits of prophylactic hysterectomy are not worth the costs (77).

Other benefits are more difficult to assess, such as the value of hysterectomy for contraception, reduction of the fear of cancer, or the elimination of unpredictable bleeding. There are no data on how many women believe hysterectomy either improved or lowered the quality of life. Even if such data were available, however, decisions about routine hysterectomy would be difficult to make. Bunker and Brown studied physicians' wives on the assumption that they would be knowledgeable consumers of medical care and found a higher rate of hysterectomy in this group than in the general population (52).

Despite these questions, the Office of Technology Assessment (OTA) has been unable to identify any clinical trial of hysterectomy underway in this country. Hysterectomy is accepted as a standard surgical procedure and reimbursed by both Medicare and Medicaid. Rates of hysterectomy vary in the United States and are associated with such factors as geographic location and type of insurance coverage. In the Medicaid program, for example, the annual rate of hysterectomy among 6,609,684 eligibles in 22 States was 303 per 100,000 population, with a range from a low of 34 per 100,000 in Mississippi to a high of 2,488 in Nevada and 1,277 in North Carolina (348).

In summary, hysterectomy is a surgical procedure that is efficacious for some conditions. But some consider it to be overused. It illustrates the difficulty of determining indications for use and of defining desirable outcomes and expected risks. Physicians and consumers appear to consider the procedure valuable. Even with the best studies, it will be difficult to make decisions concerning hysterectomy and its use (including whether Federal reimbursement programs should pay for surgery for contraceptive purposes) on fully objective bases.

Case 12: Drug Treatment for Hypertension

Hypertension, or high blood pressure, is the most common chronic disease in the United States (232). The heart generates pressure as it pumps blood to all parts of the body. Average resting blood pressure is about 120 mm of mercury systolic and 80 mm of mercury diastolic; that is, 120/80. For largely unknown reasons, this pressure can become elevated. People with high blood pressure are more likely to have strokes, heart disease, and kidney failure than people with normal blood pressure.

NHLBI estimates that 54 million people have blood pressures of 140/90 or above and require further evaluation and monitoring. At least 26 million persons have blood pressures of at least 160/95, and many of these might profit from drug therapy. At least 6.1 million persons have diastolic blood pressure above 105 mm, and all of these require drug therapy (405).

Hypertension can be effectively treated. In the late 1960's, VA carried out a multi-institutional controlled clinical trial of treatment of males for high blood pressure with the drugs hydrochlorothiazide, reserpine, and hydralazine. The control group, which was randomly selected, was given placebos. The treatment was demonstrated to be remarkably effective for men with diastolic blood pressures above 105 mm mercury. Strokes, for example, were reduced by a ratio of 4 to 1, and congestive heart failure, renal failure, and dissecting aneurysm occurred only in the control group (399). Benefits were not as clear for those with diastolic blood pressure levels below 105 mm. VA carried out an additional pilot study to collect more data on male patients with mild hypertension. NHLBI is also sponsoring further trials of men and women with all levels of hypertension, including diastolic pressure less than 105 mm mercury.

The side effects of the treatment, although seldom dangerous, are annoying. They may include dizziness, impotence, and general malaise. VA investigators state that these side effects can be minimized by careful prescription and the monitoring of treatment. Long-term use of the drugs may have side effects that are not known (60), although many of the drugs have already been in use for years.

Other questions remain unanswered. The VA study involved only relatively young male patients: does it apply equally to females; does it apply to those over age 65; what about those individuals with blood pressures under 105 mm diastolic? (126)

Furthermore, diagnosing hypertension is not easy. Validity and reliability of the measurements can be questioned for various reasons, including both systematic and random errors in reading the pressure of patients (261). Transient elevations of blood pressure are common, and care must be taken to ensure that the patient actually has hypertension (285). Many instruments for automatically determining blood pressure have been marketed; often they have not been adequately tested in the field (261).

Data obtained from national surveys based on probability samples from the early 1960's and the early 1970's indicated little change in the status of hypertension control. Approximately half of those persons with hypertension were unaware that they had elevated blood pressure and only about one-seventh had their condition adequately controlled. The VA study has led to major attempts to change this situation. NHLBI has data, collected in 1973 and 1974 from 14 communities, showing that 29 percent of hypertensives were unaware of their condition, 23 percent were aware but not undergoing therapy, 19 percent were aware but on inadequate therapy, and 29 percent were both aware and on adequate therapy. Although these data are not comparable to the national survey data, they are encouraging. In addition, patient visits for hypertension have increased dramatically in recent years (405).

The number of untreated individuals underscores the problem of "compliance," or convincing patients to take the medication. A person with hypertension must take the drugs throughout life, despite the absence of symptoms. Side effects, financial cost, and lack of explanation from physicians are some reasons that patients who feel well may not want to take prescribed drugs.

The cost of treating the entire population with diastolic blood pressures of 105 mm or greater (and a few below this level) is estimated by NHLBI at about $4.5 billion to $5 billion annually. The total cost that would be incurred if these hypertensives (those with the disease) were not treated cannot be estimated, but all cardiovascular disease, to which hypertension is a major contributor, costs society about $40 billion to $50 billion annually. Cost-benefit calculations carried out by NHLBI suggest that every dollar invested in controlling hypertension returns a benefit to society of $1.25 (405).

The Federal Government is significantly involved in the hypertension problem. FDA regulates the devices to diagnose hypertension and the drugs used to treat it. VA and NIH are sponsoring clinical trials aimed at improving knowledge. NHLBI coordinates a National High Blood Pressure Education Program, for both professionals and the public. NHLBI has also used hypertension as an example for building consensus (see chapter 5) and produced recommendations for the optimal diagnosis and treatment of hypertension for the practicing physician (285). VA has a nationwide program of screening patients for possible therapy, the Department of Defense (DOD) provides screening and therapy, and Medicare and Medicaid reimburse for treatment for hypertension, except that Medicare does not cover drugs for outpatients. Despite these efforts, a large number of patients with severe hypertension remains inadequately treated. Hypertensives are found especially in low-income groups, and blacks constitute a disproportionately large number of the individuals not being adequately treated (96).

In summary, drug treatment for hypertension has been subjected to a well-designed study for efficacy. On balance, such treatment is clearly indicated for approximately 6.1 million citizens with diastolic pressures above 104 mm mercury. It may be indicated, depending on the individual situation, for a significant portion of the estimated 20 million additional persons with blood pressures at or above 160/95, Calculations indicate that such treatment would probably be cost-beneficial. Nonetheless, despite considerable Federal activity and good efficacy and safety information, many affected individuals are not adequately treated.

Case 13: Drug Treatment for Otitis Media in Children*

Otitis media is the technical term for infection of the middle ear, a small cavity connecting the throat and the sinuses behind the ear that is necessary for effective hearing. Otitis media is believed to begin when bacteria enter the middle ear from the throat. Multiplication of these bacteria attracts white blood cells into the cavity, forming pus. The pus may burst through the eardrum and extend into the sinuses behind the ear or into the skull. Fluid can also collect in the middle ear and decrease hearing. If this fluid and the attendant loss of hearing persist, children can suffer delayed language development and impaired learning.

Ear infections are common in children. In a prospective study of 246 infants, approximately one-third were found to have ear infections at least once during the first year of life. Nineteen (8 percent) had two infections in the first year, and 4 percent had three or more infections in the first year (167). By the age of 6, 76 to 95 percent of children have had at least one ear infection. About 20 to 26 percent of children will have experienced six or more episodes by that age (172).

A variety of treatments is used for otitis media. Antibiotics are usually prescribed. Frequently, a medication for pain and a decongestant or an antihistamine are also suggested. Occasionally, a myringotomy, a simple surgical operation in which the eardrum is cut to release pus from the middle ear, is done. In about 40 percent of children, fluid persists after recovery from the acute infection (317). In these cases, antihistamines and decongestants are often prescribed and tubes are sometimes placed in the middle ear cavity through an eardrum form for draining.

Although antibiotics are accepted as efficacious therapy for ear infections, they have not been fully evaluated. They came into widespread use without careful testing about 20

*This case is adapted from material prepared for OTA by Philip Brunell, M. D., a member of the advisory panel for the study.

years ago. Controlled clinical trials to demonstrate the general efficacy of antibiotics for acute infection have been done only recently (127,317). Howie and coworkers carried out a controlled clinical trial in which the control group was given a placebo. Persistence of the middle ear infection occurred in all 45 cases of otitis caused by Pneumococcus and in 12 of 21 cases due to *Haemophilus* influenza when treated with a placebo; the most effective antibiotics cured more than 95 percent of similarly studied patients (172).

Antibiotics are also used prophylactically in children with recurrent otitis media. When Perrin, et al., tested sulfonamides in a group of children up to the age of 8 they found that prophylactic sulfonamides reduced the rate of otitis media by 7 times, with little morbidity. While sulfonamides are cheaper than most other antibiotics that might be used for prophylaxis, their nondiscriminate widespread use could be expensive for the medical care system.

The role of antibiotics in preventing the complications of otitis media is not known. Though it is difficult to find data showing a reduction in pyogenic (from pus) complications (317), most authorities agree that antibiotic therapy has decreased the incidence of acute mastoiditis, chronic eardrum perforation, and chronic mastoiditis.

The few trials of widely used decongestants and antihistamines have not shown these drugs to be effective in preventing serious otitis media (207).

FDA regulates all the drugs used for safety and efficacy. Government and private health insurance programs that include coverage for children routinely cover antibiotic treatment for otitis media as a benefit, and sometimes cover the other drugs as well. Special programs have been established for population groups with high rates of complications from otitis media, such as American Indians.

In summary, antibiotics are universally used in otitis media. After years of use, controlled clinical trials confirmed their efficacy. It appears, however, that clinical experience was adequate to demonstrate efficacy in this case, and one may question the ethics of using a placebo in studying treatments for this disease. A controlled clinical trial of prophylactic use of sulfonamides demonstrated efficacy, yet more expensive antibiotics are often prescribed. Other drugs, especially decongestants and antihistamines suggested by physicians and readily obtained over the counter in pharmacies, have no demonstrated efficacy.

Case 14: Cast Application for Forearm Fracture

Some bones, such as those in the forearm, are often fractured. Usually, the broken ends of the bone stay close to each other and, if immobilized, will heal in a period of weeks. If the ends are not close together, they are forcibly adjusted, often under anesthesia. Surgical “open” reduction with fixation by pins or other materials is also often used, despite the risk of infection or delayed healing. Experience indicates that without support during the healing process bones may not heal properly (32,156,427).

Through the centuries, various methods have been used to provide the necessary support for the bone. Ancient Egyptians, for example, used stiffened linen in a splint. The use of gypsum (plaster of paris) was first reported in 1798. Early attempts were plagued with complications such as pressure sores and gangrene caused by tight casting, stiff joints and wasting of the muscles. Techniques improved and by 1918 Bohler had developed methods still largely in use today (246).

Cast application for forearm fracture is a common procedure in medical practice. More than 1 million patient visits to office-based physicians in 1973 were for forearm

fracture, according to data from the National Ambulatory Medical Care Survey (373). Forearm fracture is the most common fracture in that study. Cast application has not been subjected to a controlled clinical trial. It is generally accepted as quite efficacious without such evaluation.

Alternatives to cast application exist, however. Traditional Chinese medicine uses different techniques. Instead of being forcibly reduced or alined, the bone ends are gradually brought into alinement, day by day. Bamboo splints are used and replaced every day. Movement of the limb begins as soon as satisfactory reduction is achieved. Horn has noted strengths and weaknesses of this method, especially its lack of complications, and described how modern and traditional methods are being merged in China (170).

Plaster of paris cast materials are regulated by FDA as medical devices. No federally supported research on cast application seems to be underway. All Government medical care programs and medical care reimbursement programs include cast application for forearm fracture as a benefit. Estimates for the annual cost of this procedure are not available.

In summary, cast application for forearm fracture is a technology whose efficacy has been established by experience in medical settings. It illustrates a technology whose efficacy could be called "manifest," that is, whose efficacy and safety are obvious to the observer. Although alternatives to cast application might be as efficacious, its widespread acceptance in this country makes development and testing of other methods unlikely and probably unnecessary.

Case 15: Treatment of Hodgkin's Disease

Hodgkin's disease, the most common neoplasm of young adults in the United States, is a form of cancer that primarily affects the lymphatic system. In 1977 there were an estimated 7,400 new cases of, and 2,900 deaths from, this disease (8).

Treatment of Hodgkin's disease primarily consists of two methods: supervoltage X-ray radiation and a four-drug combination treatment (vincristine, procarbazine, prednisone, and nitrogen mustard) known as MOPP (89). Supervoltage X-ray treatment is used for early and more localized stages of the disease and MOPP treatment for more advanced stages, although combinations of the two treatments are sometimes used.

The 3-year survival rate for patients with Hodgkin's disease increased from 35 percent in 1940-46 to 61 percent in 1965-69. From 1969 to 1973, the 5-year survival rate reached a level of 87 percent (8). The improvement resulted from new understanding of the pathology and natural history of the disease as well as development of the treatment.

In diagnosing Hodgkin's disease, pathologists classify the disease according to the predominating type of abnormal cell growth (histologic type). Laboratory tests and diagnostic X-rays are then used to determine whether the disease is confined to one lymph node region or has spread to other parts of the body. Such tests for extent of disease are called "staging. " The development of histologic: and staging criteria allowed patients to be grouped into relatively homogeneous populations according to the type and extent of disease. Knowledge of both the histologic class and the clinical stage of the disease are essential for planning the most appropriate treatment (106). Because such knowledge also permits the conduct of controlled clinical trials that are methodologically sound, the safety and efficacy of various treatments can be compared and evaluated.

Study of supervoltage X-ray treatment began in the 1930's. Controlled clinical trials of this technology have shown that 50 percent of patients with early stages of the disease

may now survive 15 years or more (107,188,273). When more extensive radiotherapy is used for limited disease, 90 percent are alive after 10 years, and most have no evidence of disease 4 or more years after treatment (205,329).

The four-drug combination treatment was developed at NCI, and its efficacy has been studied in controlled clinical trials. After completion of this treatment, 80 percent of patients with advanced Hodgkin's disease survive 5 years or more, and 47 percent remain completely free of disease (101).

Current trials are comparing new treatments and combinations with established treatments rather than with placebos or with no treatment. Controlled clinical trials are now being funded by NIH to demonstrate whether combined X-ray and drug therapy offer better results than either method alone. Other clinical trials are examining the long-term results of existing treatments (161,247,401).

In addition to evaluating the efficacy of these treatments, clinical trials provide a careful evaluation of risks. Each treatment has risks that can themselves be lethal, such as overwhelming infection (99), bone marrow suppression, pericarditis, and pneumonitis (273). A second malignancy may develop as a result of either radiotherapy or chemotherapy. In fact, recent evidence suggests that the incidence of second malignancies may be far higher in those patients receiving both radiotherapy and chemotherapy. This higher incidence may increase the risks of the therapy relative to the benefits (252). Compared to the possible benefits of a normal life span, however, these risks are considered acceptable (3).

FDA regulates the chemotherapeutic agents used in Hodgkin's disease, and FDA's Bureau of Radiological Health regulates the X-ray equipment used in treatment. In addition, the cost of supervoltage X-ray machines is high enough to require that the institution purchasing one secure a certificate-of-need (CON) from the State health planning agency. Treatments for Hodgkin's disease have been covered by third-party payers, including Medicare and Medicaid, since they first became available. Demonstration of efficacy has thus had little, if any, effect on reimbursement. In fact, ongoing trials of drugs, which could be considered experimental, are largely funded by payments of third-party payers for health services.

In summary, the efficacy and safety of treatments for Hodgkin's disease have been well demonstrated by a series of well-designed clinical trials. Insurance funds for medical services have helped to finance testing of treatments for Hodgkin's disease. The case demonstrates that testing of efficacy and safety can depend on other technologies, such as staging techniques. Additionally, the case shows that efficacy is not absolute, but relative, and requires judgments as to benefits and risks.

Case 16: **Chemotherapy for Lung Cancer**

Chemotherapy for cancer involves introducing **a** chemical or hormonal agent into the body in order to disrupt or destroy cells. It is used most frequently when surgical removal of the cancer is impossible or unsuccessful. Between 1940 and 1950, only one-third of patients diagnosed as having lung cancer were treated. From 1960 to 1970, 75 percent were treated (97). Four treatments for lung cancer have been developed: chemotherapy, irradiation (X-ray therapy), surgery, and immunotherapy. These therapies are used both individually and in combination.

Because at least 80 percent of lung cancer is caused by cigarette smoking, it is largely a preventable disease. It is nonetheless the most common form of fatal cancer in the United States, ranking first among males and fifth among females. ACS estimated that

89,000 deaths would occur from this disease in 1977 and that 98,000 new cases would be detected, a rate 14 times higher than that of 40 years ago (8). Despite the high percentage of patients who are treated, the overall 5-year survival rate for lung cancer (8 percent of males and 10 percent for females) did not change between 1950 and 1970 (97).

Multiple clinical trials of chemotherapy have led to three general conclusions about its efficacy in treating lung cancer:

1. The rate of survival of patients treated with chemotherapy for certain types of lung cancer limited to one side of the chest is similar to that of patients treated with radiotherapy and increasingly better than that of placebo-treated patients (213,432). The average increase in longevity from chemotherapy ranges from 2 to 15 months (30,58,213);
2. For extensive lung cancer, certain types of chemotherapy increase survival approximately 2 months over a placebo-treated group (30,74); and
3. The effects of chemotherapy used in combination with other therapy are unclear (58).

Durant and his coworkers compared irradiation, chemotherapy, and their combination in treating all types of inoperable lung cancer clinically confined to the chest. They found no significant difference in mean survival among the three groups. More important, they found no evidence that immediate treatment at the time of diagnosis improved either survival or quality of life when compared to the initiation of treatment when symptoms appeared, Although the study was not double-blind, it does raise important questions concerning the treatment of lung cancer patients without symptoms, especially in view of the complications of the treatment (106).

Recent evidence, however, indicates some improvements in results. According to information furnished by NCI, 20 percent of patients with oat cell carcinoma (a form of lung cancer) limited to the thorax now survive 2 years when treated with combination chemotherapy. NCI further reports that 30 to 40 percent of patients with limited non-oat cell carcinomas have increased survival periods of 14 to 15 months, up from the former median survival of 6 months.

The risks of chemotherapy are considerable and may increase in combination treatments. Many agents affect the bone marrow by lowering the number of white blood cells and thus leaving the subject liable to serious infection and even death. Another common complication is nausea or loss of appetite, with resultant weight loss and poor physical condition. Hospitalization, which affects quality of life and adds to financial costs, is often necessary during therapy.

Both methodological and ethical issues have confounded the execution of valid and reliable clinical trials. The definition of “inoperable lung cancer” has varied from study to study. Outcome measures are difficult to define. The most frequent measures have been patient survival rates and decreasing tumor size. Patients with lung cancer, however, die from other causes, and interpretation of tumor size is complicated by noncancerous disease conditions, such as infection and emphysema (74,416). These problems are further complicated by the fact that many trials compare one chemotherapeutic agent with another, rather than with **a** placebo.

Ethical problems arise in conducting such trials. If a study begins to demonstrate less improvement or greater deterioration in the treatment group than in the control or alternate treatment group, the researcher may feel ethically obligated to stop the trial.

The estimated cost for the drug for treating one patient is from $50 to $150. Approximately 60,000 new inoperable patients were treated for lung cancer with chemotherapy in 1977. Such chemotherapy is covered under most third-party reimbursement programs, including Medicare and Medicaid. Because third-party payers fund testing of chemotherapeutic agents as cancer therapy, such trials are among the least expensive at NIH.

NCI is supporting several trials of chemotherapy for lung cancer, as is VA. Chemotherapeutic agents used for lung cancer are regulated and approved for investigational use by FDA.

In summary, chemotherapy for lung cancer has been extensively studied for efficacy and safety. Efficacy is very limited. Drugs and hormones are inherently risky. Costs are high. Methodological and ethical problems plague studies in this area. Current chemotherapy for lung cancer may be a technology being diffused inappropriately.

Case 17: Hyperbaric Oxygen Treatment for Cognitive Deficits in the Elderly*

Surveys have shown that 10.0 percent of those over 65 years of age display mild to moderate cerebral dysfunction and that 4.4 percent in that age group are seriously demented, or approximately 2.2 million Americans in the first category and about 900,000 in the latter. Life expectancy is reduced to about a third of normal for the majority of seriously demented patients. The impact of mild to moderate cerebral dysfunction is more difficult to evaluate but must be highly significant in economic, social, and personal terms.

Consequently, considerable excitement was generated in both the scientific and general community when an article appeared in 1969 in the New *England Journal of Medicine* reporting enhanced cognitive functioning in elderly, male, organic brain syndrome patients following repeated exposure to pure oxygen, under pressure, in a hyperbaric chamber (1). Up to that time there was no known effective treatment for memory loss associated with brain changes due to arteriosclerotic disease or Alzheimer's disease. This finding by Jacobs and her associates (1) was even more compelling as five control subjects exposed to an air mixture failed to show improvement initially, but did improve later when they were crossed over to oxygen.

Five published reports confirmed Jacobs' observation (2,3,6,8,9). However, only one of these studies utilized a control group. Two studies failed to replicate the original Jacobs findings (10,11). One of these used 21 experimental subjects and four control subjects (11). These authors failed to note any significant differences between the experimental and control subjects.

Thus one of the major problems in evaluating the efficacy of hyperbaric oxygen as a treatment for cognitive impairment in the elderly was the paucity of studies that employed control subjects and the small number of control subjects in those that did. One reason for investigators' reluctance to include control subjects is that the control condition is more dangerous than the experimental condition. Experimental subjects breathe pure oxygen, but control subjects breathe an air mixture containing nitrogen, with some danger of the bends if care is not taken with decompression times.

Because of the importance of the Jacobs results and the obvious need for a replication study with enough control subjects to provide an adequate test of the efficacy of hyperbaric oxygen, a collaborative study was undertaken, in 1973, between the Psycho-

*This case is adapted from material prepared for OTA by the Alcohol, Drug Abuse, and Mental Health Administration.

pharmacology Research Branch of the National Institute of Mental Health (NIMH) and the New York University Medical Center.

Subjects in the study were 40 ambulatory individuals between 60 and 85 years of age residing in the community who had documented evidence of significant memory loss. There were approximately equal numbers of male and female subjects; circulatory disturbances were cited as the possible cause of organic brain syndrome in half the cases and senile brain disease was noted for the other patients.

Simply put, the results of this study failed to sustain the view that oxygen administered under pressure improves cognitive functioning in the elderly. Efforts were also made to identify subgroups of patients for whom oxygen may be especially efficacious. Again, there was no evidence of differential treatment effects as a function of initial severity of illness, sex, or presumed evidence of cerebrovascular disease. Subjects who entered this study had well-documented evidence of memory problems but were still sufficiently intact to reside in the community and to respond meaningfully to an intelligence test and to other psychological and psychometric tests. On the basis of the findings of Jacobs et al. (1) and others (2,3,6,8), one would have expected many of these patients to show a favorable response to hyperbaric oxygen treatment. The study findings clearly indicated this was not the case.

For a variety of reasons early dissemination of these negative findings was deemed in the public interest. The Jacobs findings had been picked up by the news media, especially the more sensational press, and hyperbaric oxygen was widely touted as a cure for a variety of the infirmities of old age, in addition to memory loss. A number of hyperbaric centers in this country were offering hyperbaric oxygen as a treatment for memory loss in the elderly at substantial fees. For example, at one center the fee was $5,000 for 15 days of treatment. This was not an easy issue to resolve, as scientific findings are generally not widely disseminated prior to publication in a respected scientific journal, where lag time between receipt of a manuscript and publication generally runs a year or more. To offset this delay, it was decided to present these findings at a meeting of the American Geriatric Society and to release a statement to the press once word was received that the paper had been accepted for publication (12).

Although publication of the study findings and dissemination of the results through the press and television have not completely eliminated the practice of offering this treatment to the public, it did appear to significantly dampen enthusiasm; a number of hyperbaric centers have since stopped offering this treatment. The study findings also appear to have had some impact on health insurance carriers and on the Social Security Medicare program, which at one time had considered paying for this treatment. The insurance carriers and Medicare have since ruled that hyperbaric oxygen is not a medically accepted or effective treatment for cognitive deficits in the elderly, and they will not pay for it.

The case points out the importance of appropriate dissemination of scientific findings. Information that promises relief to suffering individuals may be disseminated quickly and extensively—perhaps exceedingly so—if testing has been inadequate. It is critical that subsequent, contradictory (but more valid) findings be given the widest and most rapid dissemination.