COMPARATIVE CLINICAL EFFICACY

OF

COMPUTED TOMOGRAPHY, ULTRASONOGRAPHY, AND NUCLEAR MEDICINE

IN ABDOMINAL APPLICATIONS - A CLINICAL TRIAL

PIERRE B. BOYLE

DEPARTMENT OF EPIDEMIOLOGY AND HEALTH
McGILL UNIVERSITY, MONTREAL

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of Ph. D. (Epid.).

AUGUST 1983

ABSTRACT

Although computed tomography has received wide acceptance and rapid diffusion, its clinical efficacy relative to that of competing technologies has not been clearly demonstrated in non-neurological applications.

Two-hundred and thirty-one consecutive in-patients investigated for suspected hepatic masses or pancreatic disorders at Royal Victoria Hospital were randomly assigned to computed tomography, nuclear medicine or ultrasonography. Each of the three technologies was assessed on various measures of diagnostic and management efficacy.

Findings suggest that computed tomography has greater diagnostic efficacy than ultrasonography in pancreatic disease, although no difference in management efficacy indicators was observed. In hepatic applications, the diagnostic efficacy of computed tomography was not different than that of nuclear medicine or ultrasound; nuclear medicine showed greater management efficacy in this group than its competitors. Implications of these findings on future policy regarding diffusion and use of computed tomography and other medical technologies are discussed.

ABREGE

Malgré que la scanographie fut rapidement acceptée et diffusée, son efficacité clinique relativement à celle de technologies alternatives n'a pas été démontrée clairement dans des applications non-neurologiques.

Deux cent trente et un patients consécutifs admis à l'hôpital Royal Victoria pour étude de masses hépatiques ou de maladies du pancréas furent assignés aléatoirement à la scanographie, à la médecine nucléaire ou à l'échographie. Chaque technologie fut évaluée selon divers indicateurs d'efficacité diagnostique et de prise en charge.

Les résultats démontrent que l'efficacité diagnostique de la scanographie est supérieure à celle de l'échographie dans les applications pancréatiques; aucune différence n'a été observée dans la prise en charge de ces patients. Dans les applications hépatiques, l'efficacité diagnostique de la scanographie n'apparait pas différente de celle de l'échographie ou de la médecine nucléaire; cette dernière semble toutefois avoir un effet plus marqué sur la prise en charge.

Les implications de ces résultats quant aux politiques sur la diffusion et l'utilisation de la scanographie et d'autres technologies médicales sont discutées.

Table of Contents

List of Tables

PREFACE

Computed tomography (CT), invented in 1971 and first marketed in 1973, has been hailed as the most important addition to the physician's diagnostic armamentarium since the discovery of X-rays in the late 1800s. Although the quality and resolution of the image it provides is impressive, the effect of this information on diagnosis and clinical management over and above that which might safely be obtained otherwise has not been clearly demonstrated.

The present study focuses on the clinical efficacy of computed tomography in comparison to that of nuclear medicine, and ultrasonography in the investigation of hepatic and pancreatic disease. All three technologies are of comparable accuracy in these applications which represent a sizeable portion of the CT caseload. This study is the first randomized trial designed to compare CT to nuclear medicine and ultrasound for specific indicators of clinical efficacy.

The successful completion of the present research would not have been possible without the sustained support of many individuals. More specifically, recognition must be given to Dr. Maurice McGregor, then Chief of Medicine, and to Dr. L.D. MacLean, Chief of Surgery at the Royal Victoria Hospital in Montreal, for their open-mindedness in not only allowing but actively supporting a randomized trial on this subject, and making it possible to be carried out under the auspices of their respective departments. Their personal involvement in this project was largely responsible for the excellent cooperation obtained from both staff physicians and senior residents. Special regards are extended to

Dr. Lawrence A. Stein, Chief of the Department of Diagnostic Radiology, to Dr. R. Lisbona, head of the division of Nuclear Medicine, and to Dr. R. Patton, of the division of Ultrasonography for their expertise on the logistical aspects of the protocol and on the initial drafts of the Diagnostic Assessment Form (D.A.F.). They and their clerical staff are to be thanked for their strict adherence to protocol requirements (particularly regarding scheduling, processing and reporting of the imaging procedures of interest) throughout the six-month study. I am further indebted to Dr. L.A. Stein for his dedication and interest in integrating me into the teaching activities of the Department of Diagnostic Radiology.

Appreciation is also extended to Dr. Peter Mylnaryck (Internal Medicine), Dr. Marvin Wexler (Surgery), and Dr. Urs Steinbrecher whose participation as clinical consultants in designing the D.A.F. was active, generous and extremely valuable. They also advised me on other sources of expertise and on relevant literature items.

Very deep appreciation is extended to Dr. Kostas Psihramis, and to Miss Sally Campbell, M.Sc., who assisted in the data collection. Their dedication and scientific rigor throughout the study contributed to the high quality of the data. Dr. Pshiramis also assisted in abstracting of patient charts and in testing of abstracting procedures.

I wish to thank Dr. Barbara McNeil (Department of Radiology, Harvard School of Medicine), Dr. David Banta (Assistant Director, Office of Technology Assessment, U.S. Congress), Dr. Alvin Feinstein (School of

Public Health, Yale University), and Dr. Tom Hutchinson (Department of Epidemiology and Health, McGill University) for their critical comments on the initial drafts of the project outline; and Dr. François Béland for statistical advice. I am indebted to Mrs. Diane Bonneau; Miss Marlene Dyck and Miss Mary McLenahan for typographical assistance.

Lastly, I extend my very warm personal appreciation to Dr. John Hoey, thesis adviser, without whose relentless support and wise counsel this project might not have been undertaken nor completed.

NOTE: Sexual gender in pronouns used in this text reflects only current usage and is in no way intended to imply anything else.

This study was supported in part, by a grant from the Fond de recherche en santé du Québec and from the National Health Research Development Program, Health and Welfare Canada.

Chapter I

INTRODUCTION

A- BACKGROUND AND RATIONALE

1- DESCRIPTION OF CT TECHNOLOGY

Computed tomography (CT) is a diagnostic X-ray technique developed in 1971 in England by Godfrey Hounsfield, which provides transaxial tomographic images of sections of the head and body. CT scanning couples a narrow beam of X-rays and an electronic detector system to a computer to examine series of parallel slices of the head or parts of the body. In most commercially available models, the X-ray source and the detector rotate around the body (or head) in the same plane. As the X-ray beam passes through the various tissues of the body at various angles, X-rays may be attenuated or relatively unaffected. The variations in attenuation thus produced are recorded by scintillation detectors. The electrical signal is amplified and the data are digitized and stored in the computer which transforms the thousands of data bits into an image which represents the spatial distribution of the attenuation of X-rays in the tissues examined. The brightness of each portion of a cross-section of the body in the final image is proportional to the degree to which it absorbs X-rays. The image is then displayed on an oscilloscopic tube or television monitor and photographed for permanent recording. While most radiologic procedures provide images of projections in which over-lying structures are superimposed, CT yields transverse sections that provide a third dimensional display of the distribution of X-ray attenuation within the body. This feature as well as the sensitivity of the signal detectors used has made possible the imaging of anatomical structures and pathologic alterations heretofore invisible by conventional radiologic techniques. Conventional X-ray techniques reliably detect differences in tissue density of approximately 5% to 10%

while CT may detect density differences as small as 0.5%, thus resolving anatomical detail heretofore unobtainable. Finally, although the use of contrast enhancement agents is widespread in CT applications, CT is not considered as an invasive procedure. It can readily be used for out-patients.

Although first applied to the investigation of neurological problems (brain or head scanner), CT technology has rapidly evolved to be applied to other anatomical areas of the body (body scanner) including the abdominal area (liver, pancreas, retroperitoneum), the pelvis, chest and extremities (Evens and Jost 1979). Technological improvements to the first generation of CT scanners have primarily centered around reduced scanning time (allowing greater productivity and reducing artifact due to respiratory motion) and three-dimensional image reconstruction capability, of particular interest in spinal cord applications.

2- DIFFUSION OF CT TECHNOLOGY: ELEMENTS OF A CONTROVERSY

Given the rather spectacular improvement in image quality and resolution provided by CT over conventional radiologic techniques and its obvious potential in radiation treatment planning and monitoring, it has been termed the most important addition to the diagnostic armamentarium since the introduction of the X-ray by Roentgen in the 1800's (Cloe 1976). It is therefore not surprising that the diffusion of CT technology was extremely rapid, particularly in the United States where it is estimated that over 1000 CT units were installed within the five years following introduction of the first machine in 1973 (Fineberg 1978). According to Michel Ter-Pogossian, a radiologist at one of the first centers to use CT in the United States:

> "The acceptance of CT by the radiologic community has been immediate, unreserved, and overwhelming, even without well documented proof of its clinical usefulness."
> (Ter-Pogossian 1976)

While the benefits of CT for brain imaging were rapidly being documented, radiologists according to Ter-Pogossian (ibid.), have also enthusiastically extrapolated the use of CT to the rest of the body. The new device could reveal an impressive variety of pathologies as well as present anatomic detail previously unobtainable in vivo. For many physicians, this constituted sufficient evidence of the value of CT scanning to make it appear an essential clinical tool.

> "At no other time in the history of radiology have we been willing to purchase on faith such costly equipment."
> (M. Ter-Pogossian 1976)

In its initial report on the health policy implications of the CT scanner, the Office of Technology Assessment (1976) recognizes that CT technology was implemented widely and rapidly in the absence of sufficient information regarding efficacy and conditions for appropriate use, particularly in the case of body CT. This was substantiated by Creditor and Garrett (1977) in a study designed to explore the relationship between the information published about CT and its rate of diffusion. Creditor and Garrett indicated that only 13 clinical papers had been published in the English-language literature by June 1975 at which time 100 CT units had already been installed or purchased in the United States. The rate of diffusion of CT units in Canada and Quebec followed a similar pattern (Jacob 1982).

This lack of information base was of particular interest in the debate over CT, largely because of the cost implications of CT technology in the health care system.

The first CT units cost about $350,000 each. But as new units appeared with whole body capability and faster scanning times, the purchase price quickly increased to a level where present-generation body scanners sell for $800,000 to $900,000 while head CT units go for approximately $350,000 with slower models available for $100,000 to $200,000 (Office of Technology Assessment 1981). Operating costs associated with CT maintenance and utilization are estimated in 1982 at over $800,000 annually (Evens 1981).

Although the use of CT technology in the field of neurodiagnosis has been shown to be a useful substitute for other risky and expensive diagnostic procedures in certain patient groups (Bahr and Hodges 1978), the net cost-saving potential of head CT is still not clearly known (Knaus and Schroeder 1977; Wortzman and Holgate 1979, Jonsson and Marke 1977). This is even more true of body CT applications.

In this context, CT became an epitomy in the controversy over rising health care costs and the dilemma over marginal benefits to be derived from costly new technological advances. Concern was expressed over the cost of medical care and the expense of CT. Clinicians recognized that very promising technologies often fell short of expectations, and that even very good technologies could be overused, but were reluctant to deny their patients the potential benefits of the new technology. Health planners acknowledged the clinical promise of CT, but greatly feared excessive proliferation and sought some rational basis on which to distribute scanners. Finally, hospital administrators did not want their institutions left behind, but were concerned about such a large investment in a device whose technological obsolescence might antedate

its depreciable life (Fineberg 1978).

B- PURPOSE OF STUDY

Much of the controversy over CT - particularly of the body - centers on its capacity to add information over and above that provided by existing methods (Abrams and McNeil 1978). As Fineberg (1978) points out:

> "Diagnosis is not an end in itself. Physicians perform tests on patients to gain information about the presence or absence of disease, to help plan treatment in cases where disease is established and to monitor the results of treatment. The effect we value in its own right is the health of patients, both the length and quality of their lives, including peace of mind. In general, medicine is directed toward the goal of improved health outcomes."

However, as Abrams and McNeil (1978) point out, remoteness of health outcome from the point of a diagnostic test lead to short-term reliance on more proximal measures of a test's efficacy.

Such proximal measures of efficacy will be used in the present study designed to compare the clinical efficacy of computed tomography (CT) to existing alternatives in specific applications. The scope of the study will be limited to two groups of diagnostic applications, liver and pancreatic diseases, which constitute a large part of the total non-neurological CT caseload (Husband 1982, Evens and Jost 1979, Abrams and McNeil 1978).

Many authors (Abrams and McNeil 1978, Fineberg 1978, Wittenberg and Ferucci 1978, Wittenberg 1980, Banta and McNeil 1978, Bell 1978) have rightly held that further documentation of CT's efficacy requires careful prospective studies in which the contribution of CT is clearly related to that of competing alternative methods and the impact of

additional diagnostic information documented. More specifically, it has been recommended that in identification of mass lesions within the liver, CT be assessed in relation to ultrasound and nuclear medicine; and that it be assessed in comparison to ultrasound in the investigation of pancreatitis and pancreatic cancer (Abrams and McNeil 1978).

This prospective randomized trial of the relative efficacy of CT in specific applications, should increase our understanding of the conditions for which CT is efficacious so that proper patient triaging can be undertaken. In the present context of rationing in the field of medical care, proper use of diagnostic and other technology is essential.

Chapter II

LITERATURE REVIEW

A- DEFINING EFFICACY

The term "efficacy", often used interchangeably with the term "effectiveness", has been defined in various ways when applied to the field of medicine and health services.

Cochrane (1971) defined efficacy as referring to "the effect of a particular medical action in altering the natural history of a particular disease for the better". Using a somewhat more refined approach, the World Health Organization (1971) defined efficacy as the

> "benefit or utility to the individual of the service, treatment regimen, drug, preventive or control measure advocated or applied."

More recently, the Office of Technology Assessment of the United States Congress (1979) defined efficacy as:

> "the probability of benefit to individuals in a defined population from a medical technology applied for a given medical problem under ideal conditions of use."

Although all of these difinitions could be considered somewhat incomplete, they indicate a certain evolution and refinement of the broad concept of efficacy as it relates to medical actions and technologies. One element common to all three quoted definitions of efficacy relates to the necessary "benefit" of the medical technology or action of interest.

However, in order to achieve or produce this benefit, a medical technology must first be shown to be technically efficacious. Technical efficacy focuses on the ability of a technology to adequately do what it is designed to do, i.e. produce an interpretable image, record and plot heartbeats, etc... (Office of Technology Assessment 1976). Once a technology's technical efficacy has been demonstrated, it is possible to

assess the clinical "benefit" resulting from its output, i.e. its clinical efficacy.

B- CONCEPTS IN CLINICAL EFFICACY ASSESSMENT - ITS RELATION TO CLINICAL DECISION-MAKING

The nature of the benefit upon which the clinical efficacy of any technology will be assessed depends not only on the nature (diagnostic, therapeutic or rehabilitative) of the action or technology of interest but also on the specific moment, in the clinical decision-making process, at which one chooses to assess efficacy. There are several levels or types of clinical efficacy defined by the decision nodes in the clinical decision-making process. The following diagram (Figure 1) illustrates a simplified model of the clinical decision-making process, adapted from the original work of McNeil and Adelstein (1976). It reflects the principal concern of the present study: to compare the clinical efficacy of a new diagnostic technology (CT) to that of competing alternatives (ultrasound and nuclear medicine).

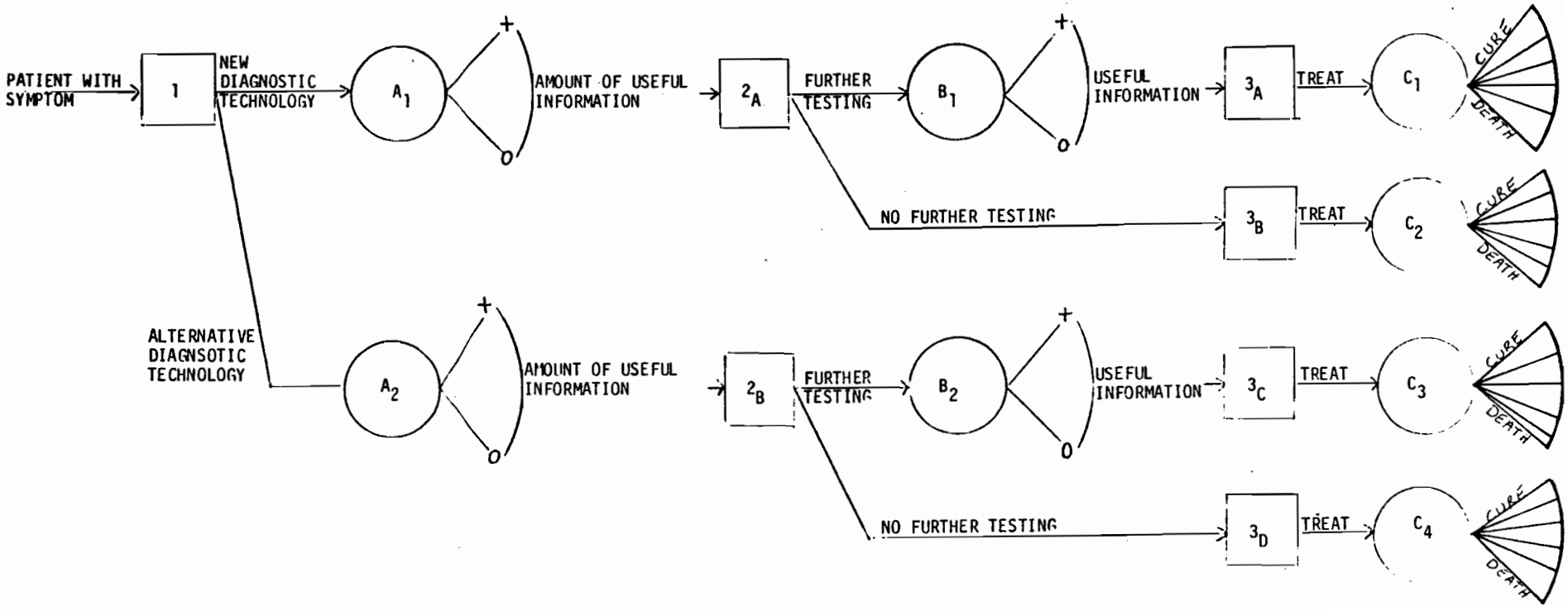
1. DIAGNOSTIC EFFICACY

In this model, a patient with a symptom complex enters the diagnostic process. Typically, the clinician will have a set of information about the patient (history, physical examination, routine laboratory tests, etc...) from which the likelihood of one or several possible diagnoses can be inferred.

At the first decision node (square 1), a new diagnostic technology or an alternative technology is chosen. In either case, the first chance node

FIGURE 1

CLINICAL DECISION-MAKING PROCESS



(circle A_1 or A_2) depicts the results of this test in terms of the amount of useful information provided. Thus, the test can provide useful information for diagnostic decision-making not previously available to the diagnostician (+) or it can provide no additional useful information (0). In the former case, the information provided may contribute to various degrees of "benefit", depending on its ability to change the clincian's understanding of the patient's condition, given all the clinical information available to the clinician at the time, as well as his knowledge of the overall accuracy of the diagnostic test in such applications.

This moment in the clinical decision-making process, illustrated by circles A_1 and A_2, constitutes the first point at which the clinical efficacy of a diagnostic procedure can be assessed, whether alone or in comparison with other competing procedures, and is termed "diagnostic efficacy" (Thornbury et al 1975; Loop and Lusted, 1978; Bell, 1978). Measures of diagnostic efficacy seek answers to the question: how much useful information is derived from the test or procedure which could not be obtained safely by other procedures? (Office of Technology Assessment 1976; Fineberg, 1977, 1; Loop and Lusted 1978; Banta and McNeil 1978). Information theory defines the "amount" of useful information provided by a diagnostic procedure on the basis of the amount of change in certainty about diagnosis, resulting from the use of a diagnostic procedure - the greater the difference between the certainty of a diagnosis after a test is performed and the level of certainty before it was performed, the greater the information content of a test (McNeil et al 1975; Barnoon and Wolfe 1972).

2. MANAGEMENT EFFICACY

Once a test's information content has been extracted, the clinician may decide either to initiate appropriate treatment at once or to request additional diagnostic procedures, to achieve greater confidence in his diagnostic assessment. This specific decision thus depends largely on the clinician's level of certainty about the nature of the patient's condition once the results of the diagnostic procedure have been reviewed and integrated with previous data by the clinician. Therefore, the diagnostic efficacy of a procedure may influence subsequent management decisions (illustrated as decision squares 2_A, 2_B and $3_{A,B,C,D}$) which constitute opportunities for assessing the "management efficacy" of the procedure. Management efficacy (Loop and Lusted 1978; Bell 1978) in the present context can simply be depicted by whether or not the management of a patient differs, as a result of the initial diagnostic procedure, from what it would have been had alternative procedures been used initially. The greater the proportion of cases where patient management differed or was altered as a result of the procedure, the greater the management efficacy of the procedure.

The importance of management efficacy was well paraphrased by Bell (1978):

> "If the handling and therapy of a patient is not changed by a diagnostic procedure, then it is likely to be regarded as useless, even if it provides a carload of information."

In summary, the basic question being addressed in assessing management efficacy in the context of competing diagnostic technologies can be formulated as follows:

> Does the information provided by the new procedure have a greater impact (than that provided by alternative techniques) on patient management decisions such as nature and duration of further testing, as well as nature, time, dose, duration and precision (particularly in oncology) of therapy?

Although some authors (Fineberg and Wittenberg 1980; Wittenberg et al 1980) have used the term "therapeutic efficacy" to describe the effect of diagnostic tests on choice, precision, dose and duration of therapy, "management efficacy" is more encompassing and better satisfies the edicts of the clinical decision-making process as it includes decision parameters related to further testing, parameters not usually included in the concept of "therapeutic efficacy".

3. OUTCOME EFFICACY

Finally, circles C_1 to C_4 (Figure 1) represent the stage at which the range of possible health outcomes may be assessed. This measure of efficacy satisfies those who contend that the ultimate test of diagnosis is the extent to which it can save lives, restore health or alleviate suffering (Banta and McNeil, 1978). In the present context, "outcome efficacy" could be defined as addressing the question:

> Does the effect of the new procedure on patient management result in better health outcomes for patients, than those resulting from the use of alternative techniques?

Typically, outcomes may vary from complete recovery to death, with several outcome gradients between those two extremes.

C- RATIONALE FOR CONCEPTUAL APPROACH TO PRESENT STUDY

In the preceeding sections, the concept of clinical efficacy was differentiated from that of technical efficacy; subsequently, three

levels of clinical efficacy were described: diagnostic efficacy, management efficacy and outcome efficacy.

The present study is not designed to address technical and outcome efficacy issues, but rather focuses exclusively on measures of diagnostic and management efficacy. The basis for this restriction will be discussed in the present section.

As outlined earlier, technical efficacy relates to a technology's ability to do what it is designed to do. In the case of diagnostic imaging technologies, technical efficacy could be defined as the ability to produce an interpretable image allowing a diagnosis to be made (Office of Technology Assessment 1976).

The technical efficacy of radioisotope and ultrasound scanning for investigation of abdominal disease has been recognized for a number of years. The wide diffusion and very rapid acceptance of the more recent CT technology (Ter-Pogossian 1976, Fagan 1977, Creditor 1977) strongly suggests that a satisfactory level of technical efficacy has been achieved. However, when comparing a new technology with a number of competing alternatives, the major concern is a comparison of the results achieved by using one modality instead of another (Banta and McNeil 1978, Bell 1978). Thus in this case, technical efficacy can be assessed as the ability of one technology to provide better images than its alternatives, allowing (more accurate) diagnosis to be made. Consequently, the relative diagnostic accuracy resulting from the use of a diagnostic technology can constitute a measure of its technical efficacy (Office of Technology Assessment 1976) relative to that of its

competing alternatives. Competing technologies for which diagnostic accuracy in similar applications has not been shown to be significantly and consistently different cannot be considered to show any difference in technical efficacy.

Although an extensive analysis and critical review of the entire literature on comparative accuracy of nuclear medicine, ultrasonography and CT will not be presented in the context of the present study, it is generally recognized that the overall accuracy of these three technologies is comparable in hepatic applications (Husband 1982, Abrams and McNeil 1978, 2; Office of Technology Assessment 1981) while CT and ultrasound are not considered to be consistently different in terms of accuracy in pancreatic applications (Abrams and McNeil 1978, 2; Office of Technology Assessment 1981).

The continued use in clinical settings, of these three technologies for the applications selected for study here, would appear to corroborate their comparable technical efficacy.

Since there is no clear basis to assume that the technical efficacy of CT is not equal to that of ultrasound or nuclear medicine in the applications under study, this study therefore focuses on the diagnostic and management efficacy of CT, ultrasound and nuclear medicine given the information available to clinicians at the time, regarding the relative comparability of their respective diagnostic accuracy.

The effect of a medical action or technology on the health status of the patient (outcome efficacy) is clearly the best approach to measuring the

"benefit" of the said action or technology. However, evaluation of outcome efficacy is difficult since health outcomes involve not only palliation, treatment and cure but should also encompass psychological well-being which clearly can be influenced by many uncontrollable variables (Banta and McNeil 1978) and take into account the value judgements made by each patient regarding each possible health outcome (McNeil et al 1977). Although such an integrated health outcome measure is surely desirable, its quantification and implementation are considered by many to be difficult (Banta & McNeil 1978; Bell 1978; Loop and Lusted 1978).

Although it has been used in some instances (George and Wagner 1975; Crighton 1962) the health outcome approach is especially problematic when applied to diagnostic medicine because frequently, health outcomes are not altered, due to the nature of the disease. For example, in those diseases for which satisfactory treatment has not yet been established (i.e. cancer of the pancreas), new diagnostic technologies, in the absence of effective treatment, cannot be expected to improve health outcomes, although diagnostic advances can in some cases predate and stimulate therapeutic progress (Abrams and McNeil, 1978, 1). Even when effective treatment is available, outcome measures are more realistically a reflection of the entire diagnostic and therapeutic process, of which diagnostic imaging is only one step (Lusted 1977).

Finally, given the slowly-developing nature of many of the conditions for which CT and its alternatives are being used in the present study and in most clinical settings, the soundest approach to assessing outcome efficacy would be based on long-term observation of patient's health

status (Abrams and McNeil 1978, 1; Bell 1978; Banta and McNeil 1978; Loop and Lusted 1978). This has been shown to be very difficult (Thornbury et al 1975) and could not be undertaken in the context of the present study.

For these reasons, the present study does not include measures of technical and outcome efficacy.

D- MEASURING DIAGNOSTIC EFFICACY

The concept of clinical efficacy outlined previously suggests that the first measure of efficacy to be considered in the present context is the amount of useful information provided by a diagnostic procedure, i.e. its diagnostic efficacy. The level of diagnostic efficacy depends on the extent to which physicians are led to change their thinking about the nature of patients' conditions, as a result of the information provided by the test (Lusted 1977).

Several measures have been used to assess the amount of useful information provided by a given diagnostic procedure.

1. YIELD

A traditional measure of efficacy in diagnostic and screening procedures is the yield or proportion of cases submitted for investigation among which an abnormal condition is discovered. This measure has been widely used in the assessment of screening programs.

In the area of diagnosis, the yield of positive examinations is an

incomplete, though not uncommon measure of efficacy (Bell and Loop 1971, George and Wagner 1975, Dronfield et al 1977). For example, the application of this measure in assessing the efficacy of cranial computed tomography was reported by French and Dublin (1977) and by Knaus and Davis (1978).

However, it must be noted that although a negative diagnostic examination has by definition zero yield, it can spare the patient unnecessary treatment (e.g. surgery), obviate the need for further examinations or better direct the subsequent diagnostic process (Loop and Lusted 1978).

Furthermore, yield is by itself an inappropriate measure of diagnostic efficacy since it is highly dependent of one factor which is independent of the technology being assessed, i.e. the prevalence of the disease among the population to whom the technology is applied.

2. ACCURACY

One obvious measure of information content is the procedure's ability to correctly sort patients with regard to specific diseases. Thus, a procedure with greater sensitivity and specificity provides more information, allowing the physician to make a correct diagnosis. The use of decision matrix and Receiver Operating Characteristic (ROC) curves have been used to assess the accuracy of a given diagnostic procedure (Adelstein et al 1970; McNeil and Adelstein 1976; McNeil et al 1975; Turner 1978; Swets 1979).

However, diagnostic efficacy, as defined here, is the ability of a test to influence a physician's understanding of a patient's condition. Since correctness of test results in any given case cannot usually be ascertained until much later in the diagnostic or therapeutic process, such knowledge cannot influence diagnostic understanding. Rather, a physician's a priori knowledge of the test's reported accuracy in similar applications becomes the baseline from which he will decide to what extent his diagnostic understanding or judgement will be influenced by the test's results.

Furthermore, the diagnostic procedures under consideration in the present study have been shown to be of comparable accuracy and thus considered to be competitive for the pathologies we have selected. Consequently accuracy is not an appropriate measure of diagnostic efficacy within the conceptual and methodological framework of the present study.

3. PREDICTABILITY OF TEST OUTCOME

Another measure of diagnostic efficacy is predictability of test outcome, an approach based on statistical decision theory (Schlaifer 1968). The theoretical concept of expected value of sample information can be applied to measure the expected value of diagnostic test information, if a medical diagnosis is regarded as a decision made in the face of uncertainty (Barnoon and Wolfe 1973; Albert 1978). The application of this framework is complex and requires the clinician to assess the conditional probabilities of different outcomes of the diagnostic procedure, given various possible correct diagnoses.

Application of a simplified version of the model in particular cases has been described by Adelstein, Parker, and Wagner (1970) and by McNeil and Adelstein (1976). Adelstein et al suggest that the information content of a test can be determined by a comparison between the predicted and measured results of the test. Thus a test, the results of which cannot be predicted correctly from the clinical manifestations of disease, would be considered efficacious. In their study, Adelstein et al compared the results of radioactive iodine uptake tests (categorized as high, low, or normal) to the probability of a normal or abnormal result given the presence or absence of 21 clinical signs and symptoms as estimated by computer analysis. On average, 21% of patients had test results containing information which could not be predicted from clinical signs and symptoms. In these patients, the RAI uptake was considered to provide additional information not obtainable from the clinical examination alone, and thus to be efficacious in those cases.

However, in order to estimate the probability of test outcome, data regarding the relative frequency of specific signs and symptoms in the presence or absence of a particular disease entity are required. Such information is rarely available. Even if it were available, the practicality of using it to design complex decision trees which physicians could use to chart test outcome probabilities is considered as dubious (Shapiro 1977, Lusted 1977). A second shortcoming refers to the insensitive nature of this measure: it does not take into account the magnitude of the difference between predicted and observed test outcomes (Shapiro 1977).

4. CHANGE IN DIAGNOSTIC LIKELIHOOD

A fourth measure of the diagnostic efficacy of a procedure is the extent to which the clinician's assessment of the likelihood of a diagnosis is modified by the information obtained from the procedure.

Since the purpose of diagnostic tests is to provide the clinician with a better understanding of the patient's condition, it seems reasonable to consider the requesting physician's diagnostic impression prior to such a test as a standard against which to measure diagnostic usefulness in any given case (Thornbury, Fryback and Edwards 1975).

Typically, the clinician's medical knowledge and data about the particular patient allow him to assign a probability for the presence of each diagnosis which is suggested by all the evidence available at the time. Then a diagnostic procedure is performed, the results obtained, added to the existing information, and new (though not necessarily different) probabilities are assigned to each of the diagnoses considered. Change in prior vs posterior diagnostic probabilities is then used to measure the test's ability to influence diagnostic judgement. Measuring change in diagnostic likelihood before and after the diagnostic procedure of interest has the advantage of pinpointing the contribution of the procedure (Bell 1978).

Several methods have been used to measure change in diagnostic likelihood.

One method of measuring change in diagnostic likelihood is to calculate the difference between posterior and prior diagnostic probability estimates as recorded by the respondent on an "open" (percentage) probability scale (Loop and Lusted 1978).

A somewhat different way of expressing certainty is with odds rather than probabilities. Information theory states that the use of odds has the advantage of eliminating scaling effects; that is the amount of information necessary to shift the odds by a given factor anywhere on the scale is the same (Barnoon and Wolfe 1972).

The use of percentages and odds in assessing diagnostic efficacy has been described and reported by Edwards (1965), Thornbury et al (1975), and more recently by Loop and Lusted 1978). These authors have focused on the application of Bayes' theorem for revising probabilities in the light of new information. The theorem can be exemplified as follows: If a physician's odds favoring the diagnosis (Dx) prior to the particular test result (T) is O_B, the Bayes' theorem states that in the case of a positive test result, his posterior diagnostic odds would be:

$$O_A = L \times O_B$$

where O_A = odds after test result (1)

O_B = odds before test result

$$L = \text{likelihood ratio} = \frac{P\ (T+ \mid Dx+)}{P\ (T+ \mid Dx-)} \quad (2)$$

One method of measuring change in diagnostic likelihood is to calculate the difference between posterior and prior diagnostic probability estimates as recorded by the respondent on an "open" (percentage) probability scale (Loop and Lusted 1978).

A somewhat different way of expressing certainty is with odds rather than probabilities. Information theory states that the use of odds has the advantage of eliminating scaling effects; that is the amount of information necessary to shift the odds by a given factor anywhere on the scale is the same (Barnoon and Wolfe 1972).

The use of percentages and odds in assessing diagnostic efficacy has been described and reported by Edwards (1965), Thornbury et al (1975), and more recently by Loop and Lusted 1978). These authors have focused on the application of Bayes' theorem for revising probabilities in the light of new information. The theorem can be exemplified as follows: If a physician's odds favoring the diagnosis (Dx) prior to the particular test result (T) is O_B, the Bayes' theorem states that in the case of a positive test result, his posterior diagnostic odds would be:

$$O_A = L \times O_B$$

where O_A = odds after test result (1)

O_B = odds before test result

$$L = \text{likelihood ratio} = \frac{P\ (T + \mid Dx +)}{P\ (T + \mid Dx -)} \quad (2)$$

and where $P(T+ \mid Dx+)$: probability of positive test result (T +) given that disease x is present (D_x +)

$P(T+ \mid Dx-)$: probability of positive test result (T +) given that disease x is absent (D_x -)

A similar equation could apply to negative test results.

But since equation 2 assumes that data on the relative frequencies of positive and of negative test results in the presence and in the absence of disease are available (which is a rare occurrence) and since the terms of the equation are meaningful independently of whether or not they refer to actual relative frequencies, likelihood ratios are usually estimated by simply converting equation 1 as follows:

$$L = \frac{O_A}{O_B}$$

Thus, the greater the difference between posterior and prior odds, the greater the likelihood ratio.

Those authors who have measured diagnostic likelihood by using probability estimates (P%) instead of odds have then converted them to odds as follows:

$$ODDS = \frac{P\%}{100 - P\%}$$

In order to obtain an additive scale for comparison of information value, the natural logarithms of individual likelihood ratios are used.

Then: $\log L = \log O_A - \log O_B$

However, each method presents some disadvantages.

a) THE USE OF ODDS METHOD

It is highly questionable whether odds reflect the thought process used by physicians in arriving at a certain judgement about diagnosis. In the 1974 - 1977 Diagnostic Efficacy Study sponsored by the American College of Radiology (Lusted 1977) respondents had to be trained extensively to record their diagnostic certainties in the form of odds. This suggests that the physician does not normally think about the likelihood of a diagnosis in terms of odds. Secondly, given the opportunity to record certainty in the form of odds or percentages, and despite extensive training, only a minority of physicians chose odds (20-30% of respondents).

The use of odds in the assessment of such a widely discussed technology as CT generates results which are difficult for the average physician to understand and relate to his clinical practice. This reduces the possible formative effect of such research. As Bell pointed out "talk of the frequency distributions and variance of the absolute values of the log likelihood ratios tends to turn off clinicians' ears." (Bell, 1978).

b) THE USE OF PERCENTAGES METHOD

Continuous percentage probability scales have been used to assess diagnostic likelihood, as exemplified in the American College of Radiology Diagnostic Efficacy Study (Lusted 1977) and in the study initiated by Wittenberg et al (1978) regarding the efficacy of computed body tomography. However, this technique presents five important sources of bias and error.

(1) SCALING EFFECTS

As mentioned previously, the use of subjective probability estimates on an open scale can lead to scaling effects (Barnoon and Wolfe, 1972) i.e. where the amount of information necessary to shift the diagnostic likelihood is different at different points on the scale. Although this problem is usually resolved by transforming percentage probabilities to odds, this transformation leads us to the problems inherent to the use of odds as described earlier.

(2) END OF SCALE EFFECTS

Data published from the American College of Radiology study (Loop and Lusted 1978) show that the initial likelihood of diagnosis was greater than 98% or less than 2% in more than 13 percent of all cases suggesting, as Bell (1978) pointed out, that "the clinician wants to be even more certain than the diagnosis initially shows." Similar results were reported by Wittenberg (1978) in his assessment of the clinical efficacy of CT.

Furthermore, despite strict instructions to the contrary, several clinicians in the American College of Radiology study jumped directly to probabilities of zero or unity. As the authors of the report recognize, this situation may not provide a true reflection of reality: "If one thinks carefully about probability assessments in the light of decision theory and practical experience he will soon recognize that probability assessments of zero or unity are seldom sensible. Moreover, clinicians who write them down almost always really meant zero simply as a very small probability and unity as a very large one."

Since the vast majority of the American College of Radiology respondents used at least the range for .01 to .99 and that fewer than 5% of clinicians were limited to the range .05 to .95, end of scale effects were of major concern to the authors. No one has yet shown whether, in fact, there is any real difference in the clinician's mind between likelihood estimates of 100%, 98% or 95% and similarly at the other end of the scale; nor at which points on the scale extremes do these effects become important.

(3) RELEVANCE TO ACTUAL THOUGHT PROCESS

There is as yet no evidence to support the notion that subjective probability assessment on a continuous scale reflects adequately the usual thought process involved in diagnostic decision-making.

The American College of Radiology study data show a heavy reliance of respondants on certain discrete probability levels (such as .01, .05, .10, .20 etc...). Such "simple values" accounted for 98.8% of all responses. As the authors conclude:

> "Although this is not in accord with much of the theoretical writing on probability assessment, we see no intrinsic reason why such rough-and-ready simplification should not serve adequately for practical application of Bayesian ideas." (Loop and Lusted 1978)

It should be noted that of the nine percent of questionnaires which were considered not usable, at least "some of them were ascribable to apparent conceptual problems with the idea of probability assessments." (Lusted 1977).

In conclusion, there is as yet no evidence that subjective probability assessments on a continuous scale constitute a method with which clinicians are readily comfortable and which corresponds to their usual thought process in diagnostic decision-making. In fact, the American College of Radiology data suggests that clinicians might be more comfortable with discrete probability intervals in assessing diagnostic likelihood and in that respect concur with the view of others (Pauker and Kassirer 1975).

(4) VALIDITY OF CONTINUOUS PROBABILITY ESTIMATES

There is little data concerning the validity of subjective probability estimates recorded on a continuous scale. However, data from the American College of Radiology study indicate that out of 197 cases of extremity injury, where the clinician thought there was a 50% probability of fracture, a fracture was present less than 36% of the time.

(5) HEURISTICS AND BIASES

Finally, certain heuristics and biases arise from making probability estimates under conditions of uncertainty (Tversky and Kahneman 1974).

i) ILLUSION OF VALIDITY

Individuals often predict the outcome (i.e. a certain diagnostic probability) that is most representative of the input (i.e. history, physical, laboratory tests) with little or no regard for the factors that limit predictive accuracy (data scanty or of questionable validity). Thus, unwarranted confidence produced by a good fit between

predicted outcome and the input information may create an "illusion of validity" (Tversky and Kahneman 1974).

ii) AVAILABILITY

Some individuals will assess the probability of an event by the ease with which similar occurrences can be brought to mind. However, availability or recall is affected by factors other than frequency; probability statements are thus subject to biases. For example: large classes are recalled better and faster than instances of less frequent classes; likely occurrences are easier to imagine than unlikely ones; associate connections between events are strengthened when the events frequently co-occur (Tversky and Kahneman 1974).

iii) ADJUSTMENT AND ANCHORING

Individuals may make estimates by starting from an initial value that is adjusted to yield the final answer, the initial value suggested by the formulation of the problem. These adjustments are typically insufficient. Thus, different starting points yield different estimates which are biased toward the initial values, a phenomenon called anchoring (Tversky and Kahneman 1974). It has also been shown that in subjective probability distributions, subjects tend to state overly narrow confidence intervals which reflect more certainty than is justified by their knowledge about the assessed quantities.

This phenomenon has also been observed in medical diagnosis. In a study of diagnostic efficacy of radioisotope scans at the Peter Bent Brigham Hospital, clinicians' posterior diagnostic probability

estimates tended to be "anchored" toward the initial values despite strong radioisotopic evidence suggesting otherwise (McNeil, B.J., 1978).

However, Tversky and Kahneman (1974) and Hogarth (1975) underline the fact that these phenomena are less important among substantive experts (such as is the case here) than among naive subjects. Furthermore, response mode i.e. using probability interval format instead of continuous probability scale, can also have a role in reducing the effects of such heuristics and biases (Hogarth 1975).

5. CONCLUSION

A review of current measures of diagnostic efficacy (yield, accuracy, predictability of test outcome, and diagnostic probability) suggests that pre-test versus post-test change in diagnostic probability is the measure most consistent with the concept of diagnostic efficacy.

Diagnostic probability or certainty can be measured by the use of odds or percentages. Although odds can eliminate scaling effects often associated with assessing changes in probability estimates, their use appears to be much less popular among physicians. This suggests that odds may be somewhat remote from clinicians' actual thought process in diagnosis.

The use of percentages in measuring diagnostic certainty or probability can be structured either on an open or continuous probability scale, or in discrete probability intervals. The latter approach is more consistent with the actual diagnostic assessment process used by

physicians, and less vulnerable to possible bias and heuristics often associated with subjective probability estimates on a continuous scale, i.e. scaling effects, end of scale effects, validity, availability and anchoring. In fact, requiring physicians to record a specific probability estimate for the presence of a particular disease may not be coherent with the state of medical knowledge.

> "At present, it may generally be said that specific probabilities are rarely known; medical diagnostic textbooks rarely give numerical values, although they may use words such as 'frequently', 'very often', or 'almost always'."
>
> (Ledley and Lusted 1959)

Although continuous probability distributions have been used for assessing diagnostic certainty, a major survey conducted by the American College of Radiology reports that physicians will overwhelmingly use "simple" numbers such as .05, .10, .20, etc... suggesting that such rounded-off probabilities are closer to actual diagnostic decision-making. Information theory also dictates great caution in subjective probability assessments. In a major review of information theory issues related to subjective probability assessments, Hogarth (1970) discusses the relevance of eliciting subjective probability distributions in these terms:

> "Man, as a selective step-wise information processing system with limited capacity is ill-suited to the task of assessing probability distributions within the framework of the more common statistical models."

Finally, there is recognition by decision analysts and information theoreticians that clinical decisions such as diagnostic assessment, further testing, choice and time of therapy are typically made and altered within a range of variation of probabilities (Pauker and Kassirer 1975). Some even argue that

> "...the purpose of prababilistic considerations (in roentgen diagnosis) is to determine which of the alternative disease complexes is 'most likely' for a particular patient" (Lusted 1960).

Eliciting likelihood estimates within certain ranges of probability corresponding to specific likelihood levels, provides a method of measuring change in diagnostic certainty which is acceptable to and understandable by clinicians, seems to reflect their actual thought processes, and minimizes the effect of certain heuristics and biases inherent to quantitative analysis of subjective probability estimates.

E- MEASURING MANAGEMENT EFFICACY

The second dimension of clinical efficacy is management efficacy which relates diagnostic information to decisions made regarding subsequent patient management.

Management efficacy assumes that a diagnostic procedure shown to improve the clinician's diagnostic understanding may affect his decision to order additional diagnostic tests (i.e. to postpone initiation of active treatment until the results of more tests are available), and may also affect the choice, dose, duration and precision of therapy. If a diagnostic procedure is not shown to change the diagnostic impression or understanding of the clinician, then clearly the procedure would not be expected to change the management of the patient and management efficacy would be zero. However, a procedure might significantly increase certainty about a diagnosis (high diagnostic efficacy) but still have zero management efficacy if there is no effective treatment for the pathology observed or no treatment alternatives. Thus no matter what the degree of certainty about the presence of the diagnosis, the

treatment would be the same.

1. FURTHER TESTING

According to the model of diagnostic and therapeutic process outlined earlier in this chapter (page 9), the first measure of management efficacy concerns the physician's decision to request other diagnostic procedures once the results of the original diagnostic procedure are known. In this context, the greater the amount of useful information provided, the lesser the need for further diagnostic tests on a given patient. In situations where competing procedures are being assessed, one can assume that the proportion of patients undergoing additional procedures will be less for the group first investigated by the most efficacious procedure than among patients investigated by the less efficacious one.

The issue of further testing has been studied in an indirect fashion by describing the total number of diagnostic procedures performed on patients with a given condition before and after the introduction of a new diagnostic modality. Knaus and Schroeder (1977) comparing aggregate institutional data of one university hospital before and after the introduction of cranial CT found an overall decrease of 14.5% in conventional radionuclide brain scanning, a 15.6% decline in cerebral arteriograms and an 80% decline in the use of pneumoencephalograms subsequent to the introduction of this new neurodiagnostic procedure. Matching hospitalized patients for diagnosis revealed significant reductions in the use of cerebral arteriography and in the use of EEG and brain scans (among the tumor group).

In a similar retrospective study of the effect of cranial CT on the use of other tests, Bahr and Hodges (1978) reported a significant decrease in the use of invasive procedures among patients investigated after the introduction of cranial CT than among pre-CT patients matched for age, diagnosis and physician. Other retrospective studies of aggregate institutional data have also been reported by Larson and Omenn (1977), Ambrose (1976), Wortzman and Holgate (1976), Johnsson and Markē (1977) and Evens and Jost (1977).

However these retrospective studies use data not generated for the purpose of the study and often include groups of patients not necessarily comparable in terms of nature and severity of diagnosis.

In addition, the introduction of cranial CT occurred at a time when many institutions were already substituting some of these procedures (Abrams and McNeil 1978, 1). Furthermore, these studies did not examine other variables which might influence the relative frequency of some diagnostic procedures (such as cost constraints, patients refusing to be submitted to "invasive" procedures, etc...).

Others (Robbins 1978, 1980; Bartlett and Neil-Dwyer, 1978) have recorded physicians' personal opinions of the extent to which, in retrospect, the new diagnostic procedure (CT) was felt to have had affected initial diagnostic plans. The very subjective nature of this approach, and the sources of error inherent to an "a posteriori" assessment without defined criteria greatly erode the validity of this approach.

Another way to measure the relative effect of diagnostic technologies on further testing, uses clinicians' diagnostic plans at the time the procedure of interest is ordered as a basis against which to measure change in patterns of additional test ordering, either prospectively or retrospectively.

Using a prospective approach, Fineberg et al (1977) and Wortzman, Holgate and Morgan (1975) asked physicians requesting a cranial CT scan: "What are the diagnostic tests you would definitely require and those you would probably require if no CT scanner were available?" A similar approach has been used by Wittenberg et al (1978 and 1980) in the assessment of body CT.

This might appear to be a hypothetical question. However, it is important to note that the reported studies were carried out relatively soon after the introduction of CT in participating research institutions; consequently it may be argued that since long-standing CT practice patterns had not been established, the use of diagnostic procedures in the absence of CT should not be considered hypothetical.

Despite such argument, the validity of diagnostic plans themselves may be questionned. When Wortzman & Holgate (1979) did a followup to their 1976 study, they observed that the reduction in the use of other neurodiagnostic procedures, projected on the basis of changes in diagnostic plans reported by clinicians as a result of CT scanning, had not materialized. The authors point to the insufficient availability and improper scheduling of CT as one of the principal causes for the difference between projected and observed reductions in the use of

non-CT neurodiagnostic procedures. Although this explanation does seem plausible (substitution is possible only to the extent that the new technology is adequately available) it does underline the fact that diagnostic plans are just "plans" and that they may differ from actual decisions because of unforeseen or changing circumstances. In fact, it may be possible that the unexpectedly smaller decrease in the use of CT alternatives was due to an unexpectedly larger proportion of cases in which other diagnostic procedures were performed after the CT scan.

On this basis it would thus appear preferable to use the actual frequency of further testing when comparing the relative management efficacy of competing diagnostic procedures. Such a measure was used by Dixon et al (1981) in a clinical trial comparing the efficacy of body CT to that of other diagnostic procedures in patients with palpable abdominal masses. Patients in the CT group had a slightly lower number of supplementary imaging investigations than patients in the non-CT group, although there was no significant difference between groups as to the number of invasive procedures. Dixon also used laparotomy as a variable to be included in the measure of further testing, such surgical procedures being by definition used for diagnostic purposes. It should be noted that all measures of further testing must refer to the same provisional diagnosis as that for which the initial procedure was performed. Otherwise a procedure, the results of which may cause the clinician to explore a new diagnostic possibility not previously suspected, would be considered as non-efficaceous if, in relative terms, it resulted in a greater frequency of additional tests (for a different diagnosis).

2. TIME TAKEN TO DIAGNOSIS

A second measure of management efficacy relates to the period of time (usually in days) taken by clinicians to make a "final" diagnosis once the diagnostic procedure of interest has been performed. The assumption is that the more efficacious a diagnostic procedure is, the more quickly a diagnosis may be made, such that therapy can be initiated.

Dixon (1981), Larson & Omenn (1977) and Evens et al (1977) have used this measure when assessing the efficacy of cranial and body CT. Levitt et al (1977) suggested this measure as one of several proposed in the framework of a computerized model for determining the diagnostic efficacy of body CT.

However, in some cases a highly efficacious procedure revealing an unsuspected diagnosis (thus causing a large change in diagnostic likelihood) may lead to subsequent confirmatory tests and thus prolong the time taken to diagnosis. It is thus essential that time taken to diagnosis refers only to the provisional diagnosis for which the initial procedure was performed.

3. EFFECT ON THERAPY

A third measure of management efficacy concerns the extent to which the information provided by a diagnostic procedure alters the choice (nature), dose, duration or precision of therapy in a given patient. Thus, a procedure would be considered efficacious to the extent that physicians are lead to treatment actions that differ from those taken in

the absence of the diagnostic procedure of interest or as a result of alternative procedures.

Three methods have been used to assess the effect of diagnostic procedures on choice or precision of therapy. The first uses changes in therapeutic plans as the measure of management efficacy and requires knowledge of whether or not the clinician would have prescribed the same therapeutic plan if the procedure had not been available. The method is based on physicians' reported intentions regarding treatment. In his assessment of cranial CT, Fineberg (1977) asked physicans to record the treatment plan that would be undertaken if the physician had to proceed without the results of a CT scan, and then compared the reported plans to the actual treatment given with the knowledge of the CT report. He concluded that any difference between the planned course of treatment and the actual treatment was a direct effect of CT. Wittenberg (1978, 1980) used this approach in assessing the efficacy of body CT.

However, there are serious pitfalls associated with this method. As Thornbury et al (1975) report, clinicians often balk at the prospect of formulating a treatment plan for a patient who has not yet undergone the diagnostic procedure. In fact, one might question the validity and logic of asking physicians to commit themselves to a certain treatment plan without the benefit of all the diagnostic information on which they would usually base their diagnostic impression. Because of these difficulties, the American College of Radiology in its radiology efficacy study (Lusted 1977) decided not to assess the effect of radiological procedures on choice of therapy. Furthermore, the use of this measure

is conditional on the availability of more than one treatment alternative, which is not always the case.

A second approach to measuring change in choice of therapy was used by Robbins et al (1978, 1980) and Baker & Way (1978) who, in assessing body CT asked clinicians or reviewers to record the extent to which, in retrospect, they felt that the choice (or precision) of therapy had been altered or improved as a result of the CT scan. The weaknesses of this subjective approach, which has also been used to measure the effect of CT on further testing, were described earlier in this section.

Finally, comparison of the use of various treatment modalities before and after the introduction of a diagnostic technology has been used to measure change in choice of therapy.

Attempting to assess the impact of cranial CT on therapy, Ambrose (1976) retrospectively compared aggregate data on the number of exploratory craniotomies performed on patients with head trauma before and after the introduction of CT in a hospital. Larson and Omenn (1977) also used this method in assessing the effect of CT on the care of 80 patients suspected of brain tumor. As indicated earlier it is often difficult in such retrospective studies, using aggregate hospital data to establish a direct or causal relationship between the therapeutic changes observed and the diagnostic procedure of interest.

Finally, a fourth approach to measuring the effect of a diagnostic procedure on choice or precision of therapy lies in clinical trials where groups of comparable patients (those with and those without the

benefit of the diagnostic procedure of interest) are compared on the basis of the therapy administered to each. Although this approach has not yet been used in the assessment of body CT, it has been used by Crighton (1962) in the assessment of cephalo-pelvimetry. He found, among patients randomly assigned to either cephalo-pelvimetry or no cephalo-pelvimetry, that patients who did have the test also had a greatly increased rate of caesarean section. Similarly, Dronfield (1976) compared operation rates of two groups of patients with upper gastrointestinal bleeding: one group was submitted to endoscopy, the other was not.

It should be pointed out that change in choice or precision of therapy is an efficacy measure which cannot be applied in the assessment of all diagnostic procedures or technologies. Crighton's application to cephalo-pelvimetry is rather straightforward in that only one kind of diagnosis could be made by this technology. At the time, this technology was the only one which could provide an accurate measurement of pelvic width against fetal head width; whether or not the test was performed, no additional tests could substitute or be complementary to it; and finally, the test results could influence only one type of therapy, i.e. whether or not to perform a caesarean-section. Thus, this diagnostic procedure could be described as being close to the decision node regarding therapeutic decisions, given the relatively short and simple clinical decision-making process. It was thus relatively easy and valid to relate therapeutic decisions to the nature of the diagnostic procedure.

However, the use of this measure is of questionable validity in

applications where the results of a diagnostic technology can, in a given case, suggest any one of several diagnostic possibilities (each likely to lead to a specific therapeutic choice). In such applications too many other contributing variables intervene between the performance of the diagnostic procedure and the decision regarding therapy to allow a direct relationship to be established. Choice and precision of therapy would thus be inappropriate for an evaluation of body CT in abdominal diagnosis.

F- METHODOLOGICAL AND MEASUREMENT ISSUES: SUMMARY AND CONCLUSION

In the present section, various issues and options concerning measures of diagnostic and management efficacy have been discussed.

Evaluation of these measures suggests that change in diagnostic likelihood recorded on a discrete scale of probability intervals is the most appropriate measure of diagnostic efficacy of the technologies under assessment here. Management efficacy can most validly be described by measures of further testing and by measures of time taken to diagnosis. Changes in choice of therapy cannot be used in the present context.

G- CLINICAL EFFICACY OF COMPUTED TOMOGRAPHY APPLIED TO ABDOMINAL DISEASE

Since computed body tomography is still a rather recent diagnostic imaging innovation, and given the stringent requirements of rigorous evaluation, few studies pertaining to the clinical efficacy of abdominal CT have yet been reported, though the literature on absolute and relative accuracy of CT is abundant.

Of the five major efficacy studies reported, three are based on physicians' personal judgement of the value of CT in a given application or group of applications. Only two studies have attempted to compare CT efficacy to that of other diagnostic imaging modalities. Only one of these consisted of a randomized clinical trial. These five studies will now be described and critically reviewed.

1. COTTON ET AL (1978)

Cotton (1978) compared the efficacy of CT to that of endoscopic retrograde cholangiopancreatography (ERCP) and grey-scale ultrasonography among 50 patients investigated for pancreatic disease and submitted to all three types of procedures. Final diagnosis was made by the investigating panel after the patient's death or discharge from hospital, based on all information available at time of analysis.

Each report for each patient was assessed by the panel at the end of the study and retrospectively given a score on the following scale:

2: correct and clinically helpful
1: correct but not clinically helpful
0: technical failure
-1: wrong but not seriously misleading
-2: wrong and clinically hazardous

With a maximum of 100, ERCP scored 75, CT 63 and ultrasound 36. Grey-scale ultrasound proved inferior to its two competitors even when technical failures were excluded, particularly in detecting the normal pancreas and providing guidance in the management of pancreatitis cases. CT scored higher than ultrasound and ERCP among patients with a known pancreatic lesion where the scans were used to differentiate cancer from

pancreatitis, although only ERCP can provide histological proof of cancer. Overall, ERCP had greater clinical impact than ultrasound or CT scanning. However, as the authors point out, results presented constituted a "crude value judgement of the clinical impact of the three imaging methods according to confidence and accuracy of the answer provided to the relevant clinical question."

Several possible sources of error and bias can be identified. One of the two elements on which efficacy was assessed was accuracy; unfortunately, data presented include patients for whom final diagnoses were made by the investigators themselves based on available information. Secondly, since the number of subjects in each of the three groups of patients is small, statistical significance may be weak - no test of significance was reported. Finally the second measure of efficacy, clinical helpfulness of the information provided by the test, was made retrospectively by members of the investigating team.

The authors failed to clearly explain the basis on which "clinical helpfulness" was assessed. Was the information clinically helpful compared to baseline data, compared to that provided by the competing modalities under study, or did it refer to the relative amount of detail provided by the test compared to that supplied by the procedure upon which the final diagnosis was based?

Finally, such studies assessing the clinical efficacy of a diagnostic procedure in a given case on the basis of diagnostic accuracy in the said case can be considered to be remote from actual clinical decision-making and an inadequate measure of clinical efficacy.

The highly subjective nature of the value ratings, the absence of any information on pre-scan diagnostic assessments and patient management plans or of any control group, as well as the absence of disease-specific efficacy scores make interpretation difficult.

2. BAKER AND WAY (1978)

Similary, Baker and Way (1978) analyzed the case histories of 202 randomly selected patients hospitalized at the University of California and Veterans' Administration Hospitals, San Francisco, who had one or more CT scans, using a 22-second scanner between January 1976 and November 1977. Details of the clinical findings, provisional diagnoses, diagnostic and therapeutic plans, and the results of relevant diagnostic tests were extracted from the chart and recorded.

Clinical efficacy of CT was measured in terms of how the results of the CT examination, exclusive of other tests improved the accuracy of the diagnosis, obviated (actually or theoretically) the need for more complex diagnostic procedures, altered treatment and affected the outcome of the disease. The authors devised an arbitrary rating scale of efficacy and each proceeded to rate retrospectively each CT scan on its overall impact on the patient's illness. If the results of the scan were judged to have had a negative effect on patient management, the extent of the negative effect was recorded on a similar scale. When independent assessments by each author were analyzed, a mean difference of only 0.4 points on the 18-point scale was observed.

Results indicate that in 169 cases (84% of total cases) the CT scan was judged to have either very little value (that is, verified a diagnosis obvious from other tests), no value, or a negative value. In 67 patients (33%) the CT scan provided information not available from other examinations. In 33 cases (16%), the results of the CT scan were used in the clinical management of patients, in particular in radiotherapy and chemotherapy planning.

In 43% of cases, CT studies were judged superfluous because simpler tests had already fully ellucidated the problem.

The authors further report that in 41% of cases the CT scan was ordered before less expensive tests (particularly ultrasound) which might have given the relevant diagnostic information. Thus, poor screening criteria for CT might explain the rather low CT efficacy ratings. Finally, in 12 patients (6%) the results of CT were considered to have had a truly negative effect on management and in three of these cases (2.0%) the patient's health was seriously threatened in the process.

Distribution of Efficacy Ratings

Rating	No. Patients	
Saved life	1 (0,5%)	
Quite valuable	4 (2%)	
Moderate value	28 (14%)	
Minimal value	118 (58%)	
No value	39 (19%)	
Obscured diagnosis	9 (4%)	169 (84%)
Seriously threatened patient's health	3 (2%)	
Led to Patient's death	0	
Total	202	

Thus, when an arbitrary rating scale was applied retrospectively to 202 in-patients submitted to CT scans, in order to assess their clinical efficacy both in terms of effect on diagnosis and management, 84% of the scans were judged to be of minimal value, no value, or negative value.

However, limitations of the study dictate caution in drawing definitive conclusions. Eligibility criteria were poorly defined. This could constitute an important source of bias, as illustrated here:

- the authors report that in 43% of cases, CT scans were superfluous since less expensive tests had already provided the relevant diagnostic information, thus reducing the potential marginal contribution of CT;

- in other cases, CT was used as a screening procedure; this might contribute to bias results towards a lower proportion of clinically efficacious CT scans;

- also included in the study population were patients who had more than one CT examination; however, there is no mention of whether or not "repeat scans" were excluded from the cases reported; if repeat scans were included, results may have been biased positively to CT efficacy (Robbins et al 1980).

Furthermore, the highly arbitrary nature of the efficacy measures used here, the absence of formal criteria to guide the value judgements which constitute the basis of the efficacy rating scale, the absence of

a control group and the absence of "blind" reviewers greatly limit the validity and generalizability of these results.

Finally, efficacy scores were not analyzed on a diagnosis-specific basis. Since Baker and Way's own data show that CT accuracy varies according to the nature of the disease investigated and since other authors (Wittenberg 1978, 1980; Robbins 1978, 1980) have reported large disease-specific variations of various efficacy measures of CT, the overall efficacy ratings reported by Baker and Way could have been quite different, had the diagnostic profile of the study population been different. Such an approach to efficacy assessment seems of little value in identifying the conditions for which CT is most or least efficacious and does not address the issue of the efficacy of CT relative to that of alternative imaging modalities.

3. WITTENBERG ET AL (1978, 1980)

During the first nineteen months of CT use at the Massachussetts General Hospital in Boston, the contribution of CT to clinicians' diagnostic understanding in selected clinical applications, its effect on choice and precision of therapy and its impact on the use of more invasive imaging tests was assessed; CT was also compared to other imaging techniques in terms of patient comfort.

Patients referred from MGH staff physicians and other practitioners were accepted for study and assigned to one of ten protocols, according to eligibility requirements specific to each protocol (diagnostic) group. The 238 referring physicians involved were urged to obtain

appropriate imaging examinations related to the protocol and were then asked to complete a pre-CT questionnaire. Of the 889 patients entered over the 19-month period, 828 were correctly assigned and had a technically satisfactory scan. A complete data base was available for 75% of the 828 patients.

The questionnaire consisted of three parts: 1- a list of possible disease complexes appropriate for the protocol, with a request to estimate the probability of each diagnosis (on a continuous probability scale from 0 to 100%); 2- a list of laboratory, endoscopic and imaging tests was provided for an opinion as to the likelihood of their performance had CT not been available; and 3- the physician was asked to indicate the provisional treatment plan based on existing information.

The scans, interpreted by three radiologists aware of results of other imaging tests, were reported as probabilities of alternative diagnoses on a list identical to that on the referring physician's first questionnaire. A handwritten report was delivered to the referring physician within 24 hours.

Upon receipt of the handwritten report, the referring physician was asked to complete a second questionnaire indicating any revision in his initial diagnostic probability estimates and therapeutic plans. If there were changes in either he was asked to indicate the extent to which, in his opinion, they depended on CT results. Diagnostic and therapeutic care was continued routinely until discharge. After patient discharge all pertinent diagnostic tests and administered

therapies were recorded. At the time of the patient's discharge from hospital the referring physician received a final questionnaire in which he was asked to record his final diagnostic impression, using the same differential diagnosis list as in the initial questionnaire. He was also asked to record his impressions of the overall usefulness of CT information in terms of identifying the diagnosis and choosing treatment modality according to the following format:

- CT DIAGNOSTIC EFFICACY RATINGS

D1 Confused understanding and led to additional tests.

D2 Confused understanding but did not lead to additional tests.

D3 Had no effect or little effect on diagnostic understanding

D4 Substantially improved diagnostic understanding

D5 Provided unique information (unavailable from any other noninvasive test)

- CT THERAPEUTIC EFFICACY RATINGS

T1 Led to treatment not in the best interests of the patient

T2 Had no effect on treatment

T3 Increased confidence in previously chosen treatment

T4 Contributed, along with other factors, to change in treatment

T5 Was largely responsible for change in treatment

Follow-up on the health status of all patients whose therapy was affected by CT was determined at least five months after examination either by chart review or by interrogation of the referring physician.

From this rather extensive set of data, the authors reported only the subjective assessments recorded by physicians on the third questionnaire.

Results indicate that for each protocol except mediastinum, the diagnostic contribution of CT was judged on the average to be greater than a small effect (D3) but less than a significant improvement (D4); CT impaired diagnostic understanding in 8% of cases, had little or no effect on diagnostic understanding in 43% of cases, and substantially improved understanding in 45% of cases. When analyzed in sequential form over the 19-month period, results indicate a significant increase over time in the relative frequency of improved diagnostic understanding for almost all protocols. Frequency of high diagnostic efficacy ratings (D4-D5) was greatest in the mediastinum, perirenal-retroperitoneum, pancreas, lung and liver protocols.

Partial results of the same study reported by Fineberg (1980) indicate that among a subset of 73 patients in whom a definitive diagnosis of pancreatic cancer, inflammation, or normal pancreas was made (based on autopsy, biopsy, or direct surgical observation), physicians' diagnostic estimates were improved in a total of 30 cases and were mislead in six cases based on the radiologist's estimate(s). One thus concludes that physicians' diagnostic probability estimates were changed (toward or away from true diagnosis) in only 36 (49%) of the pancreatic cases. However, the author did not specify the minimal magnitude of change in probability (in %) used to determine whether diagnostic assessment was altered.

Insofar as effect on therapy is concerned, CT had a detrimental effect or no effect on choice of therapy in 35% of all cases in all protocols, increased confidence in previously chosen treatment in 51% of cases, and contributed to a change in choice of therapy in 14% of cases (where surgery was the most commonly altered therapy). Impact of CT on choice of therapy was greatest for the mediastinum, liver and pancreas protocols. Over time, there was no significant change in the frequency with which choice of therapy was beneficially altered. Finally, improved diagnostic understanding (D4-D5) led to a change in choice of treatment in 29% of cases with D4-D5.

When precision of planned therapy was considered generally, and missing cases included in the "no effect" category, an improvement in the precision of planned therapy was reported for 23% of total cases. When only cases with complete data were considered, precision of therapy was improved by CT in 32% of cases in the liver protocol, and in 17% of cases in the pancreas protocol.

Data were also reported for CT's effect on planned performance of other diagnostic procedures, physicians being asked, prior to the CT procedure, to record contemplated additional diagnostic tests.

Endoscopic retrograde cholangiopancreatography (ERCP) was considered a possible procedure in 223 patients on the pancreas and jaundice protocols, and was initially contemplated in 65 of these, 44 of whom did not undergo ERCP following CT results. Unplanned ERCP was carried out, subsequent to CT examination, among four other patients.

Similarly, planned angiography was not done, subsequent to CT examination, in 83% of 76 patients for whom it was planned and was prompted in 3% of patients for whom angiography was not provisionally planned.

In summary, CT was reported to have negative, minimal or no effect on diagnostic understanding in 51% of cases while improving diagnostic understanding in 45% of cases. It contributed to a change in planned therapy in 14% of cases and to improvement of the precision of previously planned therapy in an additional 23%. CT is reported to have contributed to avoiding 68% of planned ERCP examinations and 83% of planned angiograms.

However, some caution is needed in interpreting and generalizing these results. Foremost, it is unfortunate that from the large collection of data gathered the authors reported only the results of physicians' opinions regarding the efficacy of CT from a list of proposed statements. No data is presented to describe changes in diagnostic likelihood estimates obtained from questionnaires one and two (before and after the procedure).

Recall bias may have been introduced by the fact that the subjective questionnaire was administered to physicians only at the time of discharge for hospitalized patients and several weeks after the scan, in the case of most outpatients. It is known that assessment of events remote in time tend to be influenced by more recent similar events which thus constitute the basis for judgement about more remote ones. However, it is not possible to identify the direction or effect of the

availability and recall bias in this study.

The authors report that 266 cases of the 889 eligible patients (30%) were not retained for analysis, for the following reasons:

46- inappropriate protocol assignment

15- technically unsatisfactory scans

122- questionnaire not returned or unclear

45- chart not available

38- unknown.

The authors offer no information as to the distribution of missing cases per protocol nor do they offer any information which might allow assessment of direction and magnitude of the possible error resulting from this 30% loss of eligible subjects. Clearly, the 15 technically unsatisfactory scans should not have been excluded since they represent a limitation in the use of any diagnostic imaging technique. As outlined by Fineberg (1977) the capacity of a diagnostic imaging procedure to produce a clear image allowing diagnostic appreciation constitutes the first level of efficacy (technical efficacy) at which a procedure can be assessed and has direct impact on efficacy assessments at subsequent levels. It should therefore be taken into account in the assessment of the clinical efficacy of such procedures.

No explanation is given regarding the 122 questionnaires which were not returned or were unclear, nor on the distribution of these cases by protocol. If such cases were largely concentrated among patients assigned to any particular protocol and if the physicians' judgements regarding CT efficacy in such cases were different from that in cases of

completed questionnaires then the efficacy results in one or some of the protocol groups could have been quite different.

Finally, as the authors point out, efficacy ratings were not based on histologic proof but on the final impression of the referring physician: "over-and underestimates of the worth of CT are possible" (Wittenberg et al 1980). Furthermore, effect of CT on other diagnostic procedures and on therapy was measured on the basis of planned diagnostic and therapeutic procedures. Physicians were asked to record a plan of therapy before all pertinent diagnostic information was made available to them and where also asked to retrospectively give their own opinion (once the patient had been discharged) about the extent to which any change in therapeutic plan was attributable to the CT information. This approach is subjective and vulnerable to respondents' personal biases.

Finally, the method used does not allow comparison of CT efficacy with that of other competing imaging methods in similar applications. As Fineberg (1980) points out in reporting data from the pancreas protocol group of the same study:

> "more important is the measurement in a single patient group, of the comparative efficacy of CT and other diagnostic modalities, such as radionuclide imaging or ultrasound."

Given the limitations and possible sources of error mentioned previously, as well as the weakness of the method used, results regarding CT efficacy cannot be considered conclusive. Furthermore, CT efficacy relative to that of other modalities was not assessed.

4. ROBBINS ET AL (1978)

Robbins et al (1978) using an arbitrary rating scale, assessed the clinical impact of 687 CT examinations of the chest and abdomen, performed over a one-year period at Boston's Veterans' Administration Hospital, using a $2\frac{1}{2}$ minute scanner. Prior abdominal ultrasound scanning was performed in 93% of cases. In general, when a satisfactory diagnostic end-point was achieved by ultrasound, CT was not considered. Each CT report was assessed retrospectively by the radiologist-in-charge on the basis of its impact on diagnosis and patient management according to the following categories:

a- CT provided information not otherwise available which changed diagnosis, prognosis, or therapy

b- CT provided information not otherwise available but without effect on diagnosis, prognosis, or therapy

c- no new or erroneous information was obtained from CT

Evaluation of the clinical impact of CT was made by a staff radiologist (one of the authors), 24-48 hours after the scan. Cases classified into category "a" were so classified only after agreement between two radiologists and a review of the case with the treating physician. Any case perceived by any reviewer to have contributed neither new information or to have had significant impact on diagnosis prognosis, or therapy was relegated to a lower category. Similarly, any case in which surgery was felt (by the reviewing radiologist) to have been obviated or in which planned surgery was felt to have been modified was so categorized with the concurrent judgements of senior

surgical staff. Abdominal CT was performed when intraabdominal or retroperitoneal masses, usually neoplastic or inflammatory were suspected.

Of the 271 abdominal cases studied, 23% fell into category "a", 31% into category "b", and 46% into category "c". Thus, in 77% of cases, abdominal CT was judged to be of minimal or no (marginal) value in patient diagnosis, prognosis or therapy. Of the 63 cases in category "a", abdominal surgery was considered to have been avoided as a result of the CT information in 18 cases (29%).

In a follow-up to their original study, Robbins et al (1980) studied the efficacy of CT using a two-second scanner in 101 thoracic examinations and 161 abdominal examinations performed over a four-month period. Major indications for abdominal CT were:

- suspected intra-abdominal abscess or other source of fever (29%)
- staging of known lymphoma (9%)
- evaluation of pancreas for suspected pancreatitis or cancer (19%)
- staging for other known neoplasm (22%)
- evaluation of other suspected or known intra-abdominal or retroperitoneal mass (15%)
- other (6%)

As in the original study, a large majority of patients had been submitted to ultrasound prior to CT, and almost all CT studies were, by definition, either ultrasound failures or felt to be inappropriate for ultrasound at the outset. Furthermore, excluded from the study were cases where CT, even with a positive finding, could not contribute to management

decisions.

Overall results from the follow-up study show that CT was felt to have provided information not otherwise available which altered patient diagnosis, prognosis, or treatment in 36% of cases (cagegory "a"). When the 34 repeat scans were excluded, category "a" comprised only 26% of total cases. CT reportedly provided information not otherwise available but without definite effect on diagnosis, prognosis, or treatment (category "b") in 44% of cases, corrected to 50% excluding repeat scans. CT provided no new or erroneous information (category "c") in a reported 20% of cases, corrected to 24% after excluding repeat scans. Thus in at least 65% of total cases, (actually 76% if repeat scans are excluded), CT provided no new information or had no effect on diagnosis, prognosis, or therapy. When this correct figure (76%) is compared to data reported in 1978 by the same group, no significant difference appears overall, despite the use of a much faster scanner in the more recent series, as well as the alleged improvement in radiologists' diagnostic skills with CT. Specifically, when 1980 data are corrected for repeat scans, as suggested by the authors, the proportion of cases from the abdominal group classified into category "a" appear identical to 1978 data: 23% of abdominal cases in 1980 compared to 26% in 1978. Surgery was considered to have been obviated by CT findings in 16 cases among the 94 abdominal and chest cases in category "a" in the 1980 series; it was modified or delayed in a further eleven. CT was felt to have contributed significantly to (radiotherapy) treatment planning or to modification of therapy portals in 20 cases.

Finally, 46 additional diagnostic procedures were considered to have been avoided by CT among the 272 patients in this series, most of which were tomography or arteriography involving patients in the chest group.

It is very difficult to draw conclusions about CT efficacy from the results of either of Robbins' two series. As the authors point out:

> "..,this information was not gathered by any rigidly predesigned protocol or questionnaire..." (Robbins et al 1978).

In fact, all data reported are based solely on value judgements made only after CT results were known (no data available on the diagnostic and therapeutic plans considered prior to CT results for comparative purposes). No criteria or standard was used to gauge the value judgements made by the research radiologists.

The absence of comparative data on planned diagnostic and therapeutic procedures (prior to CT scan) and actual procedures performed following CT findings makes the validity of the results rather tenuous. As stated by Loop and Lusted (1978):

> "Although (management efficacy) does not require knowledge of long run patient outcomes, it does require knowledge of whether or not the clinician would have prescribed the same management plan if the procedure had not been made available to him."

Furthermore, eligibility requirements for subjects are unclear and raise the possibility of error and bias since the research radiologists (authors) themselves excluded cases where a CT scan even if positive could not, in their opinion, contribute to management decisions. However, no specific information is given regarding the criteria according to which such exclusion were made. Consequently one can question the degree to which these data are based on actual CT

utilization on chest and abdominal diagnoses rather than on an arbitrarily selected subset of patients submitted to CT of the chest or abdomen. Finally, cases reported include both suspected and proven cases of pathology. Hence, it is likely that the overall frequency with which CT provided no new information was artificially greater than would have occurred, had the reported data included only those cases in which CT was used for diagnostic purposes only. Given that the distribution of nondiagnostic applications among the two groups (chest and abdomen) has not been reported, individual group efficacy ratings may have been biased, either positively or negatively in any one group.

5. DIXON ET AL (1981)

Dixon et al (1981) reported the first randomized clinical trial comparing efficacy of abdominal CT to that of conventional imaging modalities. The study population consisted of 60 consecutive patients with a palpable abdominal mass at clinical examination which warranted imaging. No patient had received a prior imaging investigation. Clinicians were required, in each case, to state the location of the mass and to indicate their degree of certainty about its presence on a scale of 1 to 4. They also gave a provisional diagnosis. Patients were randomly assigned to one of two groups; the first group had CT as the first imaging procedure and then as required, further imaging techniques; the second group was submitted to other imaging techniques only.

At nine months follow-up, the two groups were compared on the following factors: time taken to diagnosis, need for inpatient investigation, number of inpatient days needed to reach a diagnosis, number and risks

of imaging, need for laparotomy, accuracy of imaging.

Time taken to diagnosis was recorded as the number of days from the first investigation to a final diagnosis established either by biopsy/ laparotomy or the final imaging investigation, the latter taken to be the diagnostic endpoint when it showed no lesion or unequivocally showed a lesion which did not require immediate surgery or biopsy. Overall results are based on data available for 53 patients: 28 in the CT group and 25 in the non-CT group.

Time taken to diagnosis was significantly and considerably shorter in the CT group (mean 9.6 days) than in the non-CT group (mean 18.7 days), as was the number of inpatient days needed to reach a diagnosis (mean of 5 days for CT group vs 12.4 days for non-CT patients). No significant difference between the two groups was observed as regards the need for inpatient investigation, as well as in the mean number of investigations per patient. However, once "confirmatory tests" were excluded from both groups, patients in the CT group showed a significantly lower mean number of investigations (needed to provide a confident diagnosis) than patients in the non-CT group (means of 1.14 and 2.08 investigations per patient, respectively). However, there was no significant difference in the mean number of potentially hazardous tests (i.e. intravenous urography, cholangiography, lymphangiography, arteriography and gastroscopy) performed on patients in each group, nor in the need for laparotomy.

Since cost issues are not of concern in the present study, Dixon's results on diagnostic cost per patient will not be discussed here.

The study reported by Dixon et al has a marked advantage over previously discussed research since it compares CT to conventional imaging methods as regards certain measures of efficacy, and furthermore constitutes the first randomized controlled trial on the subject. However, important elements dictate caution in inferring from these results.

Firstly, the total number of subjects in each of the two groups was too small to achieve statistical significance for anything but very large differences between the two groups. It is thus possible that statistically significant and numerically important differences between groups could have been observed on some parameters (i.e. need for inpatient investigation, number of imaging investigations, number of potentially hazardous imaging procedures and need for laparotomy) if each group had comprised a larger number of subjects.

Secondly, diagnostic efficacy measures used by Dixon et al could be described as "proxy" measures in that they depict only the consequences of CT's diagnostic efficacy rather than representing a true measure of CT's effect on diagnostic understanding which could have been measured by various methods described previously, particularly by a before-and-after comparison of clinicians' diagnostic likelihood estimates. It is thus difficult to conclude definitely from the various measures used here, since many of them could have been influenced by several other factors related to organizational and operational elements such as lag time in scheduling a procedure (in reference to time taken to diagnosis), varying individual practices as to hospitalization of a patient for a particular investigative procedure, (need for inpatient investigation

and inpatient days needed to reach a diagnosis), to mention just a few.

A third important limitation is the fact that in this series, abdominal CT was used as the first imaging procedure and its use limited to the investigation of palpable masses. Current practice in North America is to use abdominal CT in the investigation of non-palpable as well as palpable masses, but not as a screening procedure. Thus, it is reasonable to assume that results could have been different had CT not been used as a first imaging procedure and had the applications under study not been limited to "palpable masses" only.

Finally, it is difficult to compare this study with that of others since in addition to having used a different though superior method, CT was compared to a host of other modalities analyzed as one group rather than as separate groupings. Furthermore, quite surprisingly, the make and model of the CT scanner used in this study has not been reported.

6. CONCLUSION

In summarizing the results of studies reviewed in this section, CT in various applications was considered to be of clinical value in 16% of cases while having had a detrimental effect in 6% of cases (Baker and Way 1978). In abdominal applications, CT was considered to have been of value in 26% of reported cases (Robbins 1980). When diagnostic and management efficacy measures were distinguished, CT was considered to have had significant diagnostic impact in 45% of cases, (47% for liver protocol and 54% for the pancreas protocol) while having had a detrimental effect on diagnostic understanding in 8% of total cases

(8% also for pancreas protocol; no specific figures for liver protocol). When management efficacy was considered, CT was found to have had a significant impact on therapy in 14% of total cases (19% in liver cases; 18% in pancreatic cases) while having no effect or a detrimental effect on choice of therapy in 35% of total cases (Wittenberg 1980, Fineberg 1980).

When abdominal CT efficacy was compared to other imaging techniques it was considered to be, on average somewhat less clinically helpful than ERCP but substantially more helpful than ultrasound in the investigation of pancreatic disease (Cotton 1978). Dixon et al (1981) using a controlled clinical trial, observed that patients investigated by CT for palpable abdominal masses had a significantly shorter time to diagnosis, fewer inpatient days needed to reach diagnosis and fewer investigations per patient than those investigated by a variety of conventional imaging methods.

However, caution must be exercised in comparing results from these various studies largely because of important differences in patient eligibility criteria, measures of efficacy, data groupings (organ - or disease-specific) and type of equipment used ($2\frac{1}{2}$-minute, 22-second, 18-second and 2-second scanners).

Furthermore, all but one of the five reported studies used subjective judgements on CT efficacy, to construct arbitrary rating scales. All but one of the studies failed to distinguish between measures of diagnostic efficacy and management efficacy. Finally, of the only two studies which compared measures of CT efficacy to those of other

imaging modalities, only one avoided using subjective ratings and a non-experimental design. However the very small number of subjects in the study limits the statistical power on almost half of the parameters reported and thus limit the reader's ability to gain clear insight as to the relative clinical efficacy of abdominal CT.

It is consequently understandable that so many researchers (Fineberg 1980; Abrams and McNeil, 1978; Fineberg 1977; Egdahl, 1978; Levitt 1977; Wittenberg, 1980, 1; Banta and McNeil 1978) advocated that randomized controlled trials be carried out to address the issue of the clinical efficacy of CT relative to that of competing diagnostic imaging modalities.

Wittenberg and Ferrucci (1978) after having reviewed the literature on the use and usefulness of CT as opposed to other non-invasive techniques (particularly radioisotope and ultrasound scanning) concluded:

> "Although continued technological evolution (of CT) is inevitable, major advances beyond the, as yet, untried newest generation of instruments is unlikely in the near future. A critical comparative study of these techniques is therefore in order so that unproductive diagnostic testing can be avoided. The investigations must include a randomized application of these three non-invasive tests, alone and in combination, to the spectrum of commonly encountered abdominal problems."

Chapter III

METHODS

A- STUDY DESIGN

Because of the importance of the hypotheses being tested, the possible implications of results on policy regarding the future development of the technologies being assessed and the overall purpose of the present research, a randomized controlled trial was selected as the ideal method. The favorable context of the study site made a randomized controlled trial feasible.

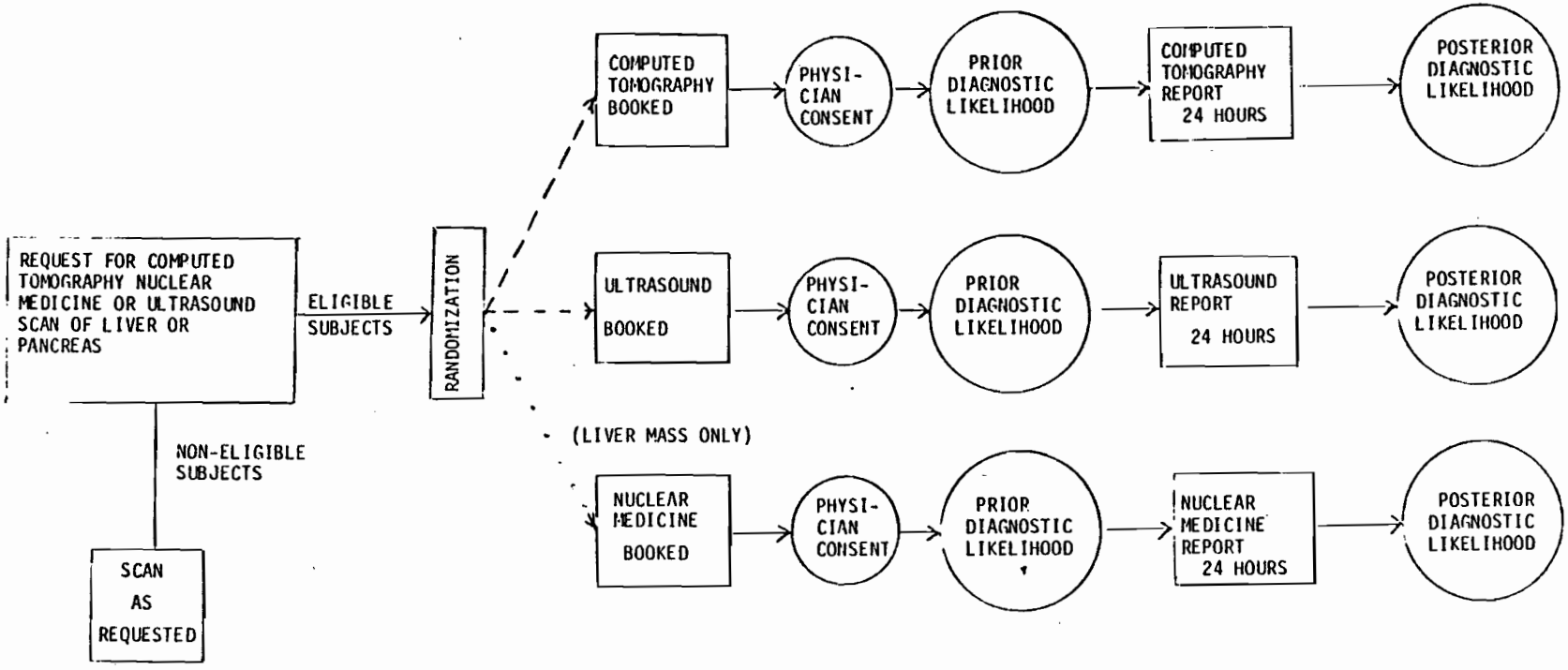
Figure no. 2 on page 64 illustrates the general design of the study, as applied to patients investigated for liver masses, jaundice, and pancreatic disease (namely pancreatitis and pancreatic cancer).

Patients to be investigated for a liver mass, jaundice or pancreatic disease and for whom a request for CT, nuclear medicine or ultrasound scan of the liver or pancreas, was received in the Department of Radiology of the Royal Victoria Hospital were screened by research staff according to the established eligibility criteria. Non-eligible patients were subsequently investigated according to the type of procedure requested by the responsible physician and according to normal department routine.

Patient characteristics (name, date of birth, sex, hospital number and service code) as imprinted on the patient's hospital admission card were recorded on a Diagnostic Assessment Form (DAF) together with the type of procedure originally requested.

FIGURE 2

PROTOCOL DESIGN



The patient was then given a random three-digit number whose last digit determined whether the patient was to be allocated to CT, nuclear medicine, or to ultrasound for scans of the liver area. But patients investigated for undiagnosed jaundice with a bilirubin level greater than 5, as well as those investigated for suspected subhepatic abscess, were randomized only to CT or to ultrasound scanning. Similarly, patients investigated for pancreatic disorders were given a random number; odd-numbered patients were assigned to CT, even-numbered patients were assigned to ultrasound. In each case, the appropriate imaging procedure was scheduled according to normal routine.

The senior resident on the ward was responsible for all patients on his ward, and is heretofore called the responsible physician. He was immediately contacted personally by the research assistant and his signed consent for the inclusion of his patient among the study subjects was secured on the DAF. The responsible physician was then asked to complete the form, recording the likelihood level (each level corresponding to a likelihood estimate interval of 25%) of each suspected diagnostic possibility among those listed on the DAF. Upon completion of the form, the physician was told the nature, date and time of the imaging procedure to which the patient had been randomly assigned. Scans were usually performed within two weekdays of the pre-scan diagnostic assessment.

Within 24 hours of the scan, a written summary report of the findings, signed by the responsible radiologist, ultrasonographer or nuclear medicine specialist, was handed by research staff to the responsible physician on the ward. In all cases, it was the same physician who had

completed the DAF prior to the scan. The physician was simultaneously handed his original pre-scan assessment on the DAF and asked to record on the same form any change in his diagnostic assessment of the patient's condition, given the scan results.

Four to six months later, trained research staff abstracted study subjects' medical records, according to a pre-determined and pre-tested abstracting procedure, and recorded on an appropriate data sheet, the nature, frequency, and date of any additional diagnostic imaging or diagnostic surgical procedure performed for the same indications, subsequent to the study scan.

B- SELECTION OF STUDY SITE AND STUDY POPULATION

1- CHOICE OF STUDY SITE

The choice of the study site depended not only on the hospital's willingness to cooperate and allow a randomized trial to be carried out but also on the satisfaction of several criteria developed in order to ensure that the trial would be conducted under optimal conditions, both professionally and technically.

The study site needed to be in a recognized university teaching hospital with a large patient caseload and a recognized clinical expertise in oncology and radiology.

The diagnostic equipment with which the study procedures were to be performed, required that the hospital have on site and in use the

following equipment:

- a second-generation computed body scanner with a maximum requirement of 20 seconds per scan;
- radionuclide imaging facilities using fourth generation scintillation camera and on-line display capabilities;
- a third-generation ultrasound scanner with grey-scale differentiation and A and B mode capabilities.

The technical and medical staff who were to perform and interpret the scan results respectively, were required to meet the following criteria:

- a certified radiologist with at least 12 months' experience in the interpretation of body CT examinations;
- a certified radiologist with special training and at least 12 months' experience in ultrasonography;
- a certified specialist in nuclear medicine, with at least 12 months' experience;
- certified radiology technicians with special training in nuclear medicine;
- certified radiology technicians with special training in the use of CT equipment.

2- THE STUDY SITE: THE ROYAL VICTORIA HOSPITAL

The Royal Victoria Hospital, incorporated in 1892 as a public hospital dedicated to treating the sick and indigent, as well as to excellence in teaching and research is the largest of McGill University's full teaching

hospitals, and one of the largest acute care facilities in Montreal.

It has a rated capacity of 743 acute care beds, 130 chronic care beds and ambulatory services in all specialties and subspecialties except pediatrics.

Clinical training programs are offered (within the context of McGill University) not only in all of its medical departments but also most other departments (nursing, physical therapy, occupational therapy, speech therapy, pharmacy, laboratory and radiological technology, biomedical engineering, social work).

The Department of Diagnostic Radiology (radiologist-in-chief, Dr. Lawrence A. Stein) is staffed with ten active members, including one member in the division of ultrasonography and one member in the division of nuclear medicine. They are supported by technical and clerical employees. The department is recognized by the Corporation Professionnelle des médecins du Québec, the Ministère de l'Education du Québec and the Royal College of Physicians and Surgeons of Canada for postgraduate training in diagnostic radiology and nuclear medicine, as well as being active in the rotating internship and undergraduate programs of the Faculty of Medicine of McGill University.

3- CHOICE OF CLINICAL PROBLEMS FOR STUDY

The primary purpose of the study was to compare the diagnostic and management efficacy of CT, nuclear medicine and ultrasonography in similar applications. Choice of the clinical applications to be

included in the study was made, therefore, on the basis of the relative comparability of the overall accuracy of the three technologies, and on the relative volume of the chosen applications.

As mentioned in Chapter II, the three diagnostic technologies under study are considered to have relatively comparable levels of overall accuracy in abdominal applications, particularly liver and pancreas, although their relative sensitivity and specificity may differ. Furthermore, abdominal applications represented a large portion of the use of these three technologies.

The choice of the actual diagnostic possibilities used in the study within the area of abdominal applications was made in cooperation with a group of clinical consultants using the two general criteria mentioned above.

4- SELECTION OF STUDY SUBJECTS

The study population comprised all hospitalized patients of the Royal Victoria Hospital for whom a written request (signed by house staff or attending staff) for a CT, nuclear medicine or ultrasound scan of the liver or pancreas was received in the Department of Diagnostic Radiology during the period from December 3, 1978 to June 5, 1979 (excluding weekends and statutory holidays). This six-month observation period was necessary in order to enter the required number of study subjects.

The following exclusion criteria were strictly followed. Excluded patients were those:

1- whose original imaging requisition had not been signed or verbally approved by a staffman or senior resident;

2- under the age of 18 (CT having been shown to constitute a possible risk for certain patients in younger age groups);

3- who had undergone investigation of the liver or pancreas by either CT, nuclear medicine or ultrasound scan in the six-month period preceeding the present request.

All other members of the study population, as defined previously, were considered as eligible study subjects. Eligible subjects who, for one of the following reasons, were not able to complete all stages of the study protocol were not replaced; they are reported separately:

1- the responsible physician refused to allow the patient's participation;

2- the patient refused to undergo the scan to which he or she had been randomly assigned;

3- the procedure was cancelled by the responsible physician because of a change (marked improvement or deterioration) in the patient's condition.

C- PLANNING THE STUDY

1- DEVELOPMENT OF STUDY DESIGN AND INSTRUMENTS

The overall study design was elaborated from a preliminary outline submitted to the thesis director in May 1978.

In June 1978, the preliminary outline was forwarded to the Physician-in-chief and to the Radiologist-in-chief of the Royal Victoria Hospital. An informal meeting with the thesis director and Radiologist-in-chief was then arranged as a means of assessing the general interest for and feasibility of such a project, within the Department of Radiology (the randomized trial put forward in the initial project outline had already been turned down by other major Montreal teaching hospitals).

During June, July and August, a more thorough literature review as well as critical comments supplied by consultants(1), within and without the McGill Faculty of Medicine provided the necessary elements to prepare a preliminary version of the proposed research protocol which was completed at the end of August 1978.

---

(1) Including: Dr. Lawrence Stein, Royal Victoria Hospital
Dr. L. Rosenthall, Montreal General Hospital
Dr. Maurice McGregor, Royal Victoria Hospital
Dr. Barbara McNeil, Harvard University, Department of Nuclear Medicine
Dr. David Banta, Office of Technology Assessment, United States Congress
Dr. Harvey Fineberg, Harvard School of Public Health
Dr. Alvan Feinstein, Yale University School of Medicine

This more comprehensive proposal was then forwarded to those who had commented on the initial project outline. It was also forwarded to the Radiologist-in-chief, to the head of ultrasonography and to the head of nuclear medicine who were invited to meet with the principal investigator and his thesis supervisor to discuss the possibility of undertaking the proposed study, under the joint auspices of McGill's Department of Epidemiology and Health, and the Royal Victoria Hospital's Department of Diagnostic Radiology.

At a meeting held on September 11, 1978 it was decided by the Radiologist-in-chief (Dr. L.A. Stein) that the Department of Diagnostic Radiology would support, promote and participate in the proposed study. At the same time, a list of methodological and logistical issues to be clarified was developed.

An estimate of the total number of cases of undiagnosed jaundice, liver masses and pancreatic masses investigated by either nuclear medicine, ultrasound or CT in the previous six months was supplied by the Department of Radiology, as well as information regarding the usual process for requesting, scheduling, performing, interpreting and reporting such scans of the liver and pancreas.

As of mid-September, the Chief of the Department of Diagnostic Radiology, the head of the Division of Nuclear Medicine, and the head of the Division of Ultrasonography accepted to serve as members of an advisory

panel of diagnosticians, set up by the principal investigator.

The mandate of the advisory group of diagnosticians was to assist in:

1- designing the required randomization protocol(s) by identifying through consensus the best (imaging) alternative to CT, for each type of possible diagnosis for which a scan of the liver or pancreas could be requested;

2- identifying appropriate subject eligibility criteria and estimating the expected monthly volume of eligible patients per diagnostic group;

3- designing a standard data collection instrument;

4- drafting a general procedure for patient screening, randomization, scheduling, pre-scan diagnostic assessment by physician, interpretation and reporting of scan results, post-scan diagnostic assessment;

5- identifying a list of diagnostic imaging procedures which could likely be ordered by the responsible physician, once the clinical findings of the study scan had been reported.

This group was also consulted regarding patient consent as well as the definition and treatment of technically poor scans. A detailed outline of the mandate accepted by the members of the advisory group of diagnosticians appears in Appendix A.

Similarly, a select group of clinicians, recommended by their respective department heads (chief of medicine and chief of surgery) accepted the invitation to serve as members of the Advisory Panel of clinicians. This group was composed of an internist specialized in gastrointestinal disease, an oncology surgeon, and a senior medical resident.

All members of both advisory groups were active senior members of the clinical (or house staff) of the Royal Victoria Hospital, as well as members of the teaching staff of McGill University Faculty of Medicine (with the exception of the senior medical resident).

The responsibilities of the advisory group of clinicians were:

- to assist the principal investigator in preparing a list of possible diagnoses for the attending physician whose patient was a study subject;

- to cooperate with the advisory group of diagnosticians in designing a general data collection procedure, in choosing subject eligibility criteria, and finalizing the necessary randomization protocol(s) including the identification of the best alternative to CT scanning in each diagnostic or protocol group;

- to determine the need for written patient consent.

A detailed outline of the group's mandate appears in Appendix A. The two groups worked separately at the beginning, then met jointly on issues such as finalizing the list of diagnostic possibilities, the

formal randomization protocols and the general data collection procedure.

On October 13, 1978, all panel members received a summary of the decisions or recommendations unanimously approved by both advisory groups concerning the relevant aspects of the study, and were asked to forward any correction or comment deemed appropriate. That document (Appendix A) served as a working basis for the overall study design, procedure and instrumentation (to be adapted and finalized in the following weeks) and became part of the research proposal which was submitted to a joint Ethics Committee of the McGill Faculty of Medicine and the Royal Victoria Hospital.

2- APPROVAL BY HOSPITAL AUTHORITIES

On November 1st, 1978, the Ethics Committee was convened under the chairmanship of the chief of the Department of Medicine of McGill University and the Royal Victoria Hospital. The committee was composed of the Director of Professional Services of the Royal Victoria Hospital, of a member of the Division of gastroenterology of McGill University and the Royal Victoria Hospital, as well as the head of the division of Nuclear Medicine of McGill University and of the Montreal General Hospital. All members of the committee were active in both clinical and academic activities.

The following issues were raised by the committee and discussed with the principal investigator of the project:

- randomization procedure
- feasibility and duration
- alternative plan of study
- patient consent
- cost implications for the hospital
- policy implications

On the issue of patient consent, the principal investigator stated that informed written consent from the patient was not required because:

1- Under normal clinical circumstances, patient's written consent is not required specifically for each diagnostic procedure performed while in hospital (the patient is required to give a general consent to all procedures deemed necessary by the attending physician);

2- on the basis of the available information, no patient could be considered as suffering a prejudice as a result of the randomization process;

3- in all cases the attending physician was responsible for selecting the appropriate diagnostic procedure or delegated this responsibility to the senior resident on the ward;

4- the protocol allowed the physician to request any supplementary diagnostic procedure as soon as results from the study procedure were available, thus informed consent by the attending physician or senior resident should be sufficient.

The Ethics Committee recommended to the Associate Dean of Medicine for Graduate Studies and Research, that the proposed study be accepted, subject to inclusion on the DAF of the following statement to be signed in each case by the responsible physician (senior resident) and the staff physician on the service:

> "Having been informed of the purpose and nature of the above-titled study, I agree to participate and hereby give consent to the random allocation of this patient to either ultrasound, computed tomography, or radionuclide scanning (the latter excepted in pancreatic evaluation). I understand that such randomization will not involve any additional risk for the patient, and that I am free to request any supplementary investigation considered necessary following review of the study scan results".

It should be noted that one month into the actual trial the staff physicians' consent was abandoned as a required trial practice. The senior resident(s) continued to be required to give consent for each case.

Upon acceptance of the proposed study by the Ethics Committee, the chiefs of the Departments of Medicine and Surgery respectively agreed to lend their full support to the study. On November 16, 1978 a letter (Appendix A) under the joint signature of the Chief of Radiology, the Chief of Medicine and the Chief of Surgery, was sent to all members of the medical and surgical staff of the Royal Victoria Hospital explaining the purpose and general design of the study, advising them of the starting date, approximate duration and potential benefits of the study, and finally urging them to lend their full support to its realization.

3- PRE-TEST

During a two-week period, from November 20 to December 1st, 1978, a pre-test of the entire randomization and data collection procedure was carried out on a total of 26 eligible patients.

The purpose of the pre-test was to:

a) verify estimates of projected weekly volume of eligible subjects;

b) validate feasibility of the data collection procedure, particularly with regard to scheduling of procedures, screening of eligible subjects, feasibility of obtaining diagnostic imaging reports within 24 hours, and physician participation;

c) determine if clerical staff in the Department of Radiology would proceed as directed and that instructions would be understood;

d) test the construct validity of the DAF, specifically evaluating two modes for registering diagnostic likelihood: either on a continuous scale from 5% to 95% probability, or on a four-point scale ranging from very unlikely (< 25% likelihood) to very likely (> 75% likelihood);

e) test the acceptability of the list of diagnostic possibilities on which the clinician was to be questioned.

As a result of the pre-test:

a) the general design of the DAF was modified to render its reading and comprehension simpler, and to allow registration of the reason for which the scan was requested;

b) the four-point scale of diagnostic likelihood was retained as the preferred method (preferred by 5 out of 7 participants in 20 of 26 cases) for recording diagnostic likelihood. The scale was constructed on the basis of a four-point rather than a five-point scale, in order to reduce central tendency bias. Because of the nature and implications of the diagnostic procedures under study, because of the importance of their rational use, and because of the importance of establishing a diagnosis, within the general context of clinical decision-making, it was decided that the scale should not include a middle-point, nor should it allow the respondent to answer: don't know. It was felt that the use of these diagnostic imaging technologies should be preceeded by baseline diagnostic information (history and physical, laboratory tests, etc...) sufficient to allow a provisional diagnosis to be made whose likelihood would be assessed by the clinician before the test procedure was performed.

c) the three participating diagnosticians interpreting all scans included in the study, were advised and agreed that a hand-written or typed scan report signed by the diagnostician was necessary for the study and that no verbal report was to be given to any physician (except in absolute emergency situation) unless the principal investigator had already conveyed the written report of the Department of Radiology to the responsible physician.

d) minor adjustments were made to the logistics of data collection.

As the pre-test was being carried out, the teaching and research staff of the Department of Epidemiology and Health (McGill University) were invited to express their critical comments on the proposed study protocol and data collection instruments during a two-hour seminar held by the investigator in cooperation with the thesis director.

Comments received on this occasion as well as suggestions made by various consultants from other research centers also served to develop the final version of the research design and instrumentation.

4- TRAINING OF HOSPITAL'S CLERICAL STAFF

One week prior to the beginning of the pre-test period, the clerical staff involved in the screening, scheduling of patients and forwarding of reports were interviewed. They were given a written summary of the goals and methods of the proposed study and their role was explained. The importance of their dedicated and sustained participation was underlined. Despite the fact that some aspects of the design involved some extra work and required their devoted attention, all four clerical staff members enthusiastically agreed to participate.

A detailed outline of the procedure they were to follow within the context of the study (Appendix B) was handed to each of them, two days before the beginning of the pre-test. On the eve of the pre-test period, each member of the clerical staff was visited by the principal investigator and his assistant, in order to review the procedure and answer any pertinent questions. The procedure to be followed by the clerical staff was posted in their respective offices, together with the names and phone numbers of both the principal investigator and his

research assistant. Subsequent to the pre-test period, a slightly modified version of the procedural instructions was given and explained to the clerical staff who in fact had suggested some changes of their own.

5- TRAINING OF THE CLINICAL STAFF

a) PRE-TEST

During the week prior to beginning of the pre-test, the staffman as well as the resident and intern staff on each of five medical and surgical wards met with the principal investigator. On that occasion, they were handed a two-page outline describing the purpose, method and duration of the study, a list of persons associated with the study, as well as copy of the DAF to be used to record pre-scan and post-scan diagnostic assessment.

The house staff were given instructions as to the completion of the DAF, as well as to the procedure to be followed when submitting a requisition for CT, nuclear medicine or ultrasound scan of the liver or pancreas. The staff on each ward was visited by the research staff occasionally during the first few days of the pre-test, in order to respond to any problems encountered.

b- THE STUDY

During the first week of the actual study, the Chief of Medicine and the Chief of Surgery allowed the principal investigator to address the medical and surgical staff respectively, at weekly rounds, usually attended by most senior staff persons and all resident staff.

After a supportive introduction by the Chief of the department, the principal investigator gave a thorough but concise presentation of the proposed study (purpose, methods, implications) and of the actual procedure to be followed by the house staff. A two-page summary of the project was distributed to those present as well as a copy of the DAF.

The staff were also informed of the starting date and estimated duration of the trial. Since the senior resident on each ward was the person who would actually complete the DAF, the principal investigator met individually with the senior resident on each of the participating wards, on the first morning of the trial, in order to make sure the procedure was understood and followed.

As part of the McGill-Royal Victoria Hospital graduate programs in the Departments of Medicine and Surgery, the resident staff on each ward is rotated every two months. On the first day of each rotation, the new resident staff on each participating ward were given the same information and training process described above.

6- TRAINING OF RESEARCH ASSISTANTS

Throughout the duration of the study, the principal investigator was supported by two research assistants each for a three-month period(1).

---

(1) Dr. K. Psihramis, M.D., served as research assistant from December 1, 1978 to March 1, 1979.

Ms. Sally Campbell, B.Sc., M.Sc. (Epid) from March 15 to June 3, 1979

The research assistants were briefed on the research issues, the purpose, scope, methods and implications of the study. All relevant documents pertaining to the study were reviewed. Each assistant was given six days practical training on data collection procedures. Data collection responsibilities were shared with the investigator.

The research assistant was introduced to all clinical and clerical staff participating in the study, as well as to all senior medical and surgical residents on the participating wards. Furthermore, the research assistant observed as the principal investigator went through the actual randomization, scheduling and data collection procedures in the hospital.

After appropriate instruction, the research assistant actually carried out those procedures under the immediate supervision of the principal investigator. All questions raised by the research assistant were clearly and immediately discussed, so that by the end of the training period, the assistant was fully capable of assuming the responsibilities of managing the day-to-day operations of the study.

The principal investigator remained available to the research assistant at all times during the course of the study and could be on site or reached by phone within minutes, should any unforeseen difficulty arise. In addition, the assistant carried at all times a "Data Collection Procedure and Daily Schedule" (Appendix B), which outlined the procedure to be followed.

D- INSTRUMENTATION - DESIGN OF DIAGNOSTIC ASSESSMENT FORM

Pre-scan and post-scan data regarding the likelihood of various diagnostic possibilities were recorded by the responsible physician on the patient's own DAF which appears as Appendix C.

The form was composed of five parts:

1) IDENTIFICATION OF THE STUDY

The top portion of the form bore the following identification: "McGill-RVH Diagnostic Efficacy Study for Suspected Hepatic and Pancreatic Masses."

2- PATIENT IDENTIFICATION

In the top right-hand corner of the form was a blank space on which was imprinted the patient's hospital card, bearing the following information: patient's name, surname, date of birth, sex, hospital number, date of admission, admitting clinical service or unit, and name of admitting staff physician.

3- REASON FOR SCAN

This section was designed to serve as a second screen for scans requested for non-diagnostic purposes (i.e. to guide or assess therapy, to assess prognosis, etc...) and to identify those scans requested to confirm and those requested to rule out a diagnosis.

In this section, the clinician was simply asked to indicate the "reason for ordering this test" - his choice was limited to one (only) of the categories listed below:

- to rule in or confirm a diagnosis
- to rule out a diagnosis
- to guide therapy of an established diagnosis
- to have basis to assess efficacy of planned therapy
- to assess prognosis.

As it turned out, none of the respondents indicated a reason other than one of the first two. This suggests that screening procedures for non-diagnostic indications used in the Department of Radiology at the time the scan request was received, were correctly applied.

4- DIAGNOSTIC ASSESSMENT

This fourth and most important section of the DAF offered the clinician a list of 20 diagnostic possibilities to choose from. For hepatic and pancreatic disease, each possibility provided a space in which the physician would record his assessment of the likelihoodof presence of the diagnosis for which the patient was being investigated. The DAF instructed the respondent physician to "select the diagnosis(es) for which this patient is being investigated, limiting your choice to no more than three diagnostic possibilities, and indicate the likelihood that the chosen diagnosis(es) is (are) present."

In fact, in only 6 of the 248 total cases did the respondent physician indicate more than one diagnostic possibility (other than "normal").

Six weeks into the study, the respondents were directed verbally to ignore "normal" as one of the diagnostic possibilities to consider. This decision was made since the possibility of a "normal" organ seemed,

from the data collected over a period of 6 (six) weeks, to apply however remotely to almost all cases. Thus, because a choice of "normal" was non-specific, did not add useful information for the purpose of the study, and because respondents felt compelled to indicate it in almost all cases, this possible category was abandoned.

The DAF instructed respondents to "rate the likelihood from 1 to 4" according to the following scale:

1- very unlikely (less than 25% probability)

2- unlikely (25-49% probability)

3- likely (50-74% probability)

4- very likely (57% or more probability).

Diagnostic possibilities for liver disease were listed in three groups: masses, undiagnosted jaundice, normal. Masses were classified into four categories: infectious, neoplastic, vascular and other. Each category then branched into various diagnostic possibilities as follows:

Masses-Infectious- Abscess - Intrahepatic
- Extrahepatic - subhepatic
- subphrenic

-Neoplastic -Primary benign
-primary malignant
-Secondary

-Vascular -Hematoma
-A-V Malformation

-Other (specify)

Jaundice NYD -Obstructive
-Non-obstructive

Normal

Similarly for pancreatic disorders, the list of diagnostic possibilities was constructed as follows:

Masses-Pancreatitis - Acute
- Chronic - with pseudocyst
- with abscess

-Neoplastic - Primary benign
- Primary malignant
- Secondary (metastatic)

-Other (specify)

5- PHYSICIAN CONSENT

This last portion of the form contained the consent statement previously described (page 77).

E- MEASURES OF DIAGNOSTIC EFFICACY

Three different measures of diagnostic efficacy were used, each measure corresponding to one of the research questions listed in Chapter II:

- the frequency of change in diagnostic likelihood observed subsequent to the scan of interest
- the magnitude of that change as measured by a diagnostic efficacy score
- improvement in diagnostic understanding

1- CHANGE IN DIAGNOSTIC LIKELIHOOD

The rationale for selecting likelihood or probability intervals constructed on the basis of a four-point scale has already been discussed. The instrument used in the present study for assessing the

physician's pre-scan and post-scan diagnostic impression has also been described.

This first measure of diagnostic efficacy aimed at providing a basis for comparison of the three technologies' respective ability to effect a change in the physician's diagnostic impression. It can be defined simply as the relative frequency (proportion) of cases investigated for a specified condition by one of the technologies, among whom the post-scan diagnostic likelihood (level) assessed immediately after scan results were reported to the clinician, differed from the pre-scan diagnostic likelihood (level).

2- MAGNITUDE OF CHANGE IN DIAGNOSTIC LIKELIHOOD

The second measure of relative diagnostic efficacy describes the importance of the change in diagnostic likelihood when it occurs. It was developed to take into account;

a) the numerical magnitude of the change on the four-point scale;

b) the possibility that changes of equal numerical importance on the likelihood scale may have very different consequences in terms of patient management and thus different levels of clinical significance for the treating physician;

c) and conversely, that changes of unequal numerical importance on the likelihood scale, might have similar clinical significance.

Thus, all possible changes on the likelihood scale were given a weight which was designed to adjust the numerical magnitude of change on the

scale, for the perceived relative clinical importance of the change for future patient management, as outlined below. This was called the clinical function value of the change in diagnostic likelihood.

The relative weights were originally assigned on an arbitrary basis. The resulting weighting system was then submitted independently to eight (8) senior clinicians for comment and adjustment. A second version was developed and submitted to six other senior clinicians, five of whom found the (new) weighting proposal to be acceptable. The weighting system was constructed around the following principles:

a) single unit changes between the two middle values on the four-point scale (2 and 3) are the least clinically important of all changes;

b) single unit changes leading to or away from the two extremes of the scale (i.e. 1 and 4) are more important clinically than those in a) since they are more likely to lead to a medical decision;

c) double or triple unit changes have the same importance as the sum of the corresponding single-unit changes required to equate them.

WEIGHTING OF CHANGES ON DIAGNOSTIC LIKELIHOOD SCALE

Numerical Changes on Likelihood Scale	Numerical Value	+	Clinical Function Value	=	Resulting Weighted Value
1 ↔ 2	1	+	1	=	2
2 ↔ 3	1	+	0	=	1
3 ↔ 4	1	+	1	=	2
1 ↔ 3	2	+	1	=	3
2 ↔ 4	2	+	1	=	3
1 ↔ 4	3	+	2	=	5

Thus, any shift in likelihood to or away from one of the extreme values on the scale would receive one "bonus point" over and beyond the numerical value of the change. For example, an initial likelihood rating of "one" (0-25% probability) which is changed to "three" (50-74%) after the procedure would receive a value of two (2) points for the change of two levels on the scale, in addition to one extra point for shifting the likelihood level away from the extremities on the scale, resulting in a weighted value of three (3).

Similarly, changes from one extremity point on the scale to the other extremity would receive two "bonus points" over and above the numerical value (of three) for having shifted the likelihood estimate by three levels (i.e. 1 to 4 or from 4 to 1), thus resulting in a weighted value of five.

The analysis reported in section E of Chapter IV focuses on a comparison of the mean weighted score of each protocol group.

3- IMPROVEMENT IN DIAGNOSTIC UNDERSTANDING

The third and final indicator of diagnostic efficacy relates to the very purpose of a diagnostic procedure, which is to contribute to the improvement of the physician's understanding of a patient's condition. The three technologies under assessment will be compared for this indicator, on the basis of two measures.

The first measure deals with the overall change (increase or decrease) within each protocol group of the proportion of cases in the very unlikely and very likely categories; the greater the increase, as a result of the scan, in the proportion of such cases, the greater the diagnostic efficacy of the technology used.

The second measure of improvement in diagnostic understanding focuses specifically on instances where a change in diagnostic likelihood was observed; such changes are then classified, in each case, as constituting, or not, an improvement in diagnostic understanding depending on direction of the change in diagnostic likelihood (i.e. either towards or away from the extreme points on the diagnostic likelihood scale).

F-MEASURES OF MANAGEMENT EFFICACY

1- FREQUENCY AND NATURE OF SUPPLEMENTARY TESTING

The first measure used to compare the management efficacy of CT, nuclear medicine and ultrasonography rests on the postulate that the greater the clinical usefulness of a diagnostic imaging procedure for a

clinician, the less likely he is to request additional diagnostic procedures. The null hypothesis is that there is no difference between the technologies as to the proportion of similar patients who were submitted to supplementary diagnostic procedure(s) of a radiological or surgical nature. Consequently, any technology showing a significantly lower proportion of patients undergoing additional tests would be considered as providing more useful diagnostic information than its competitors.

It must be pointed out that the measure developed here bears only on the relative frequency of patients submitted (or not) to one or more additional tests and not on the number of additional tests ordered (per patient) as such. It was decided that the (mean) number of additional tests per patient was not an adequate measure of management efficacy of the initial diagnostic scan, since the decision to submit a patient to a second post-test procedure is a function of the information provided by the first post-test procedure, rather than of the information provided by the initial scan. In fact, information theory and clinical decision analysis suggest that the decision to order or not to order a diagnostic test is largely influenced by the result of (and information supplied by) the immediately preceeding diagnostic procedure, rather than that of other prior tests. Consequently, the present section focuses essentially on the presence or absence of supplementary testing (for each patient) as well as on the nature of the first post-scan supplementary procedure ordered.

2- DURATION OF POST-SCAN RADIOLOGICAL AND SURGICAL INVESTIGATION

This final indicator of management efficacy stems from the concept of "time taken to arrive at a final diagnosis", described in the previous chapter. Because of the difficulty in many cases of identifying a correct diagnosis based on pathological or historical proof, the concept of time taken to diagnosis was not retained.

Instead, it was decided that a more valid measure of management efficacy would be the period of time (after the initial scan) during which a patient was investigated by diagnostic imaging or surgical means, for confirmation or exclusion of the same suspected pathology as that for which the initial scan was performed.

Accordingly, the technology whose patients had the shortest period of post-scan investigation for the same indication would have the greatest degree of management efficacy.

G- PATIENT SCREENING AND RANDOMIZATION PROCEDURE

Three series of random numbers, one each for the three protocol groups, were drawn from random number tables published in Arkin H., Colton R.R., Tables for Statisticians, Barnes and Noble, New York, 1963.

Group one was designed to include patients investigated for all liver masses (except subhepatic abscess) and those investigated for jaundice of unknown origin; subjects in this group were randomly allocated to either CT, nuclear medicine or to ultrasonography.

A list of 175 three-digit numbers between 101 and 699 (excluding those ending with a zero) were drawn consecutively without replacement from the sequential and horizontally continuous numbers from lines 1 to 25 of the fourth thousand, and lines 1 to 5 of the fifth thousand blocks of random numbers published in Arkin H. and Colton R.R. (1950), pages 159-160. All numbers were used.

Group two was designed to include only patients investigated for subhepatic abscess. Subjects in this group were randomly allocated to either CT or ultrasonography. Forty-two two-digit random numbers between 01 and 99 were generated consecutively without replacement from the sequential order of horizontal numbers on lines 2, 4, 6 and 8 of the second thousand block of random numbers published on page 158 of the aforementioned text. The prefix number "8" was added such that all study numbers for this group of patients were three-digit numbers and such that the subjects in this group not be confused with those of any other.

Group three was designed to include those patients investigated for pancreatitis or for a pancreatic tumor. Subjects in this group were randomly assigned to either CT, or ultrasonography.

A list of 67 two-digit numbers between 01 and 99 were generated consecutively without replacement from the sequential order of numbers on lines 3, 6, 9, 12 and 15 of the third thousand block of random numbers published on page 159 of the aforementioned text. The prefix number "9" was added to each two-digit number, in order that the group be correctly identified.

The following procedures were used to enter patients into the study. All requisitions for diagnostic imaging procedures to be done on in-patients at the Royal Victoria Hospital were forwarded to the Department of Diagnostic Radiology. Those requisitions concerning specifically CT, nuclear medicine and ultrasound procedures were directed to the offices of the Chief of Diagnostic Radiology, the head of Nuclear Medicine, and the head of the Ultrasonography division, respectively, usually before 11:00 a.m., Monday through Friday.

The requisition form for each of the three procedures contained information regarding patient and physician identification, patient location, type of procedure, organ to be visualized, and a brief clinical summary or at least the reason for the requested procedure.

The clerical staff in the respective offices held all requisitions for CT, nuclear medicine and ultrasound scanning of the liver and pancreas areas to be done on hospitalized patients. At regular and frequent intervals throughout the day, the research staff collected the said requisitions, verified that the patient was hospitalized, verified the nature of the scan requested and the organ to be visualized, and finally ascertained compliance with eligibility criteria by reviewing departmental files (located in the same offices and organized in alphabetical order of patients' names) for any similar scan performed on the same patient for a similar indication within the previous six months. All requisitions had to be approved in writing or verbally by the responsible staff physician or senior resident.

Upon determining that all eligibility criteria had been met, the patient's name was recorded on the log sheet for the appropriate protocol group, next to the first available random number listed on the log sheet. This number was to be the patient's study number.

The nature of the procedure to which the patient had been randomized was determined by the last digit of the study number (which had been randomly assigned to the patient) as follows:

- Liver mass and jaundice group: patients whose study number ended with 1, 2, or 3 were assigned to nuclear medicine; those whose number ended with 4, 5 or 6 were assigned to ultrasound; those with study numbers ending with 7, 8 or 9 were assigned to CT.

- Subhepatic mass group: patients with even numbers were assigned to ultrasound; those with odd numbers were assigned to CT.

- Pancreatic disease group: patients with even study numbers were assigned to ultrasound; patients with odd study numbers were assigned to CT.

As the patient's name was recorded on the appropriate protocol log sheet, the patient's identification data (full name, sex, date of birth, clinical service, date of admission, hospital number) were transcribed from the requisition onto the patient's DAF.

Also recorded on the form were the patient's study number as well as the nature of the procedure originally requested by the responsible

physician.

The original requisition was then modified, if necessary, to reflect any change in the nature of the procedure to be performed as a result of randomization. A red dot on the top right-hand corner of the requisition alerted the staff of the Radiology Department that scheduling of procedure and reporting of clinical findings for this patient were to be done according to study procedure(s).

Arrangements were immediately made in the Department of Diagnostic Radiology to schedule the patient for the procedure to which he (she) had been randomly assigned. The procedure was to be done within two working days of randomization.

H- INSTRUMENTATION AND METHODS FOR DIAGNOSTIC IMAGING & INTERPRETATION

1- COMPUTED TOMOGRAPHY

All scans of study subjects were performed on a EMI 5005, eighteen-second scanner located at the Montreal Neurological Institute, physically linked to the Royal Victoria Hospital (and its Department of Diagnostic Radiology) by a connecting corridor.

Liver scans were performed with an infusion of 75cc of Hypaque 60 used as contrast material. No oral Hypaque was given.

However, in the case of pancreatic applications, 400cc of a 2% solution of oral Hypaque was given one hour prior to the examination. No intravenous Hypaque was given for examination of the pancreas.

In almost all situations, patients were examined in the supine position. In very unusual circumstances, different views were taken. All examinations were done at 140KV 28 Ma.

Criteria for interpretation of the scans were as follows:

- Jaundice: Dilated bile ducts were reported when branching tubular structures were demonstrated near the hilum of liver, often extending down medially to the duodenum through the head of the pancreas. If filling defects were seen in the bile ducts, they were so reported. When enlargement of the liver was seen with dilated bile ducts, carcinoma had to be excluded.

- Abscess: This was usually seen as a lucent mass within the liver, with an irregular margin and often with a rim sign. It is difficult to differentiate abscess from a hepatoma although a hepatoma usually has a higher CT number.

- Metastatic disease of the liver: These are usually well defined areas of slightly decreased density within the liver.

- Pancreatitis: Usually diagnosed by enlargement of the pancreas, ductal ectasia, with or without pseudocyst formation.

- Carcinoma of the pancreas: Usually diagnosed by a localized area of enlargement in the pancreas of a slightly less CT number than the remaining pancreas. This enlargement is usually lobular in nature and is often associated with lymphadenopathy.

2- NUCLEAR MEDICINE

All nuclear medicine scans of study subjects were performed by a 1977 SEARLE PHO GAMMA V camera. The radiopharmaceutical used for liver scans consisted of sulfur colloid, labeled with 3 millicuries (mCi) of technicium 99 (^{99}Tc). The dose was administered intravenously and the patient was imaged 15 minutes after the injection.

Four views were obtained in each case: anterior, posterior, and right and left lateral projections. A total of 400,000 counts per image was accumulated in each case.

Typical reports were either:

- no evidence of space occupying lesions

or - one or more space occupying lesions are present in the liver.

3- ULTRASONOGRAPHY

Abdominal B-scanning was done with a Picker Echoview System 80-L, in conjunction with transducers namely a 2.25 mHz long focus 19mm., a 3.5 mHz long focus 19mm.; and 3.5 mHz. medium focus 13mm. Very occasionaly a 5 mHz. transducer was used.

Patients were instructed not to eat or drink for 8 to 12 hours before the procedure. This usually emptied the stomach and freed it of gas. Occasionally Simethicone in water was administered if epigastric gas was present.

Most patients were examined in the supine position and 1 cm. cuts were made in the longitudinal and transverse positions, of the area of interest. Utilizing known structures such as the aorta, inferior vena cava and gall bladder, the intensity settings were adjusted to the individual. When a mass was identified, the intensity was varied to aid in distinguishing between cystic and solid structures.

In obstruction of the biliary system, dilated ducts were reported when, in the hilum, there were many echo-free tubular structures. The dilated common ducts, with or without evidence of calculi, were followed towards the pancreatic area. These examinations were performed with the highest frequency transducer possible and with the smallest head diameter.

The liver was examined in 1 cm. cuts in longitudinal section from the lateral margin of the left lobe to the lateral margin of the right lobe. The examination was performed in deep suspended inspiration. Transverse sections were also performed to further delineate an abnormal mass.

The pancreas was usually examined in the supine position. Occasionally the examination was performed with various maneuvers such as in the left or right decubitus position, with or without water in the stomach, and with or without the addition of simethicone. The pancreatic region was identified by its relationships to the normal vascular anatomy such as the splenic vein, superior mesenteric vein, superior mesenteric artery and inferior vena cava. The cuts were made in the longitudinal, transverse and occasionally in the oblique positions.

The examinations were recorded on X-ray film utilizing a Dunn camera. The examination was reported after review of the films and in view of the clinical history.

I- PROCEDURE FOR DATA COLLECTION

1- PRE-SCAN DIAGNOSTIC ASSESSMENT

Once the randomly assigned procedure had been scheduled and preliminary (patient) information recorded on the DAF, the research staff met the senior resident on the service responsible for the patient's care, and asked him to complete the DAF (and sign the consent statement included therein), recording the reason for the requested procedure, and the likelihood of (each of) the diagnosis(es) for which the patient was being investigated.

Once the DAF had been completed and signed, the senior resident was informed of the nature (and often of the date) of the scanning procedure to which the patient had been randomized. In all cases, the senior resident was told that the date and time of the procedure would be confirmed in advance by the staff of the Department of Diagnostic Radiology. In the case of patients scheduled for abdominal ultrasonography, the senior resident was reminded that the patient should receive nothing by mouth for 8 to 12 hours prior to the procedure.

2- POST-SCAN DIAGNOSTIC ASSESSMENT

The scheduled scans were performed within the normal delays, i.e.: usually the following working day for radioisotope and ultrasound scans,

and within two days for computed body tomography. All scans were interpreted only by the Chief of Diagnostic Radiology (for CT scans), by the head of the Division of Nuclear Medicine (for nuclear medicine scans) and by the head of the Division of Ultrasonography (for ultrasound scans). The only exception to this rule was for a two-week period during which CT scans were read by a senior member of the Department of Radiology (with special training in computed tomography) in replacement of the Director of the Department, away on holidays; and a two-week period during which the head of ultrasound was away--he was replaced by a senior member of the Department of Radiology with training in ultrasonography. All examinations were performed and scans interpreted as described.

Within 24 hours of the study scan, a written report signed by the appointed diagnostician (as above) was issued by the Department of Radiology. No verbal report was to be given to the responsible physician before he had seen the written report, except in cases of emergency of which there were none. Furthermore, the clerical staff of the Department of Radiology screened any other request for diagnostic imaging of the liver or pancreas for the same patient, received after a study subject had been scheduled for a study procedure, and reported such cases to the research staff.

In such cases, the requested (supplementary) procedure was scheduled and performed, sometimes even before the study scan was done, but in all such cases, results of the (supplementary) procedure were witheld by the Department of Radiology until the post-scan DAF had been completed.

However, this raised the problem of whether, had the scan results been available earlier, the other (supplementary) diagnostic procedure would still have been requested i.e.; whether this procedure should in fact be considered as supplementary, for the purposes of the present study. This unexpected problem was identified two months and 45 cases into the study. It was decided that from this point onward, the responsible physician who had ordered another imaging procedure (but still had not obtained the results) before the study scan report had been handed to him would be asked, after completing the post-scan DAF, whether "in view of the scan findings he would maintain his request for the additional procedure". The answer was to be given verbally to the research assistant in the form of "yes" or "no". The research staff was strictly ordered not to discuss the matter with the clinician at that time nor to answer any question the clinician might have in that regard. The resulting verbal answer was recorded by the research staff in the top right-hand corner of the DAF. Seventeen cases "double-ordering" were recorded and processed in this fashion (out of a total of 186 cases) during the remaining four months of the study. All of them concerned patients investigated for a liver mass by either ultrasound (ten cases) or CT (seven cases). Twelve of the seventeen were considered supplementary, since the clinician still considered the additional procedure necessary even after having read the study scan report. Nine of the twelve cases involved patients investigated by ultrasonography; the other three were patients initially investigated by CT.

During the course of chart abstracting, it was observed that six other cases of double-ordering had occurred during the first two months of the study. Since for those cases there was no empirical basis for identifying truly supplementary procedures from the others, it was decided to attribute to them on a relative basis, the results of the information obtained for the subsequent 17 cases of double-ordering, particularly in view of the fact that the type of case was similar: four patients initially investigated by ultrasound and two patients initially investigated by CT for a suspected liver mass. Hence, four of the six cases were recorded as having had a supplementary diagnostic procedure, three of them being patients in the ultrasound liver mass group, the other patient from the CT liver mass group.

Written scan reports were collected by the research staff at regular intervals throughout the day at the appropriate office(s). The report was then immediately handed by the research staff to the responsible physician (senior resident) on the ward. Having read the report, the responsible physician was then handed the DAF on which his pre-scan diagnostic assessment had been recorded, and asked to record his present diagnostic assessment (immediately beside his initial likelihood levels) taking into account the scan results just reported. Under no circumstances did the research staff provide the responsible physician with any additional explanation.

In cases where the senior resident questioned the diagnostic findings reported, he was invited to contact the appropriate diagnostician in the Department of Diagnostic Radiology.

3- SUPPLEMENTARY DIAGNOSTIC PROCEDURES: FREQUENCY, NATURE AND DURATION

Four to six months after the end of the actual study, the staff of the Medical Records Department identified the records of all patients, according to a list of patient names and their respective hospital numbers, supplied by the principal investigator, and appearing in chronological order of entry into the study.

Charts were abstracted by the principal investigator and a research assistant according to a standard procedure (Appendix C). The abstracting procedure was validated for content and reliability by using 30 records, each being abstracted independently for the pertinent information by the principal investigator, and his assistant.

In each case, the abstracted information was recorded by each abstractor on a separate patient data sheet. Once all 30 records had been abstracted by each abstractor, the information was compared for similarity of interpretation and completeness to that recorded by the other abstractor.

Only the supplementary diagnostic procedure(s) relevant to the condition for which the patient had been initially scanned, was recorded. This was made possible by the fact that all diagnostic imaging and biopsy requisitions at the Royal Victoria Hospital routinely contain the clinical indication or a brief clinical account for the requested procedure. A few minor clarifications were brought to the original abstracting procedure, subsequent to the validation and reliability tests. All subsequent records were abstracted according to the revised

procedure, and the information recorded directly by the abstractor on an individual patient data sheet.

During abstraction, study scan results as reported in writing for each patient, and included in the patient's hospital record, were transcribed onto the patient data sheet.

4- DATA TABULATION, CODING AND ANALYSIS

Relevant information recording socio-demographic characteristics of study subjects, their study number, the nature of the initially requested scan, the nature and date of the study procedure to which the subject had been randomly allocated, the reason for the requested scan, the name of the responsible physician as well as pre-scan and post-scan diagnostic likelihoods were transcribed by the investigator from the DAF onto an individual patient data sheet (APPENDIX C). Data were verified by an assistant, for any transcription error.

Data abstracted from patient charts regarding the following elements were recorded directly by the abstractors on the patient data sheet: clinical findings of the study scan, the nature and date of supplementary diagnostic procedures, date of patient's discharge from hospital, post-scan length of hospital stay, post-scan duration of radiological and surgical investigation, primary discharge diagnosis, relevant discharge diagnosis (if different from the primary discharge diagnosis), ascertainment of discharge diagnosis and accuracy of the study scan.

These data were subsequently coded directly on the patient data sheet,

by an experienced coder, according to the coding manual appearing in APPENDIX C. Each study record was verified for coding errors by the investigator.

Data were subsequently analyzed for differences between protocol groups on specific indicators of diagnostic and management efficacy. The chi-square statistic was used in most cases where comparison of proportions were involved; comparison of means were done either with t-test or analysis of variance. When appropriate, the d statistic was used in comparing means, as suggested originally by Welch (1951) and illustrated by Armitage (1971). Log-linear analysis with the G^2 statistic was used for one indicator where between-group comparison of within-group changes over time was being analyzed.

All of the analyses, except log-linear analysis were performed with the use of a Hewlett-Packard HP41C calculator; the log-linear analysis was done on a terminal using the ANOMHI program (Béland 1980).

Note: For a detailed dscription of the meaning and use of the G^2 statistic, see: Haberman, S., Analysis of Qualitative Data, vol. 1, Academic Press, New York, 1978.

Chapter IV

DATA DESCRIPTION AND ANALYSIS

A- STUDY POPULATION AND OUTCOME OF RANDOMIZATION

During the pre-determined duration of the study (December 3, 1978 to June 5, 1979), a total of 231 patients were considered as satisfying the eligibility criteria outlined in Chapter III and all of them were subsequently put through unrestricted randomization (also outlined in Chapter III) and given a time reservation for the diagnostic imaging technology to which each had been randomly assigned.

The results of the randomization process appear in Table II, according to the diagnostic protocol groups described in Chapter III. It would seem that, generally, the randomization procedure achieved its goal of randomly assigning each patient to one of the study procedures appropriate for his (her) group, and consequently distributing the subjects in about equal numbers to each of the procedures within a group.

However, in group II (undiagnosed jaundice and liver masses except subhepatic abscess) the number of patients assigned to ULTRASOUND is unexpectedly high (64) compared to the number of patients in each of the two other procedure groups, i.e. nuclear medicine (56) and CT (53). After careful analysis, this would seem to be exclusively the result of the actual sequence of random numbers used in the process of randomization.

As described in Table I, twenty-one (21) of the 231 subjects entered into the study were lost to view. Six of those subjects died before the study scan could be performed; in five other cases, the assigned imaging procedure was cancelled by the attending physician because of a sudden and unexpected change (improvement or deterioration) in the

TABLE I

FLOW OF STUDY POPULATION

1-	PATIENTS ELIGIBLE		231
2-	PATIENTS ENTERED INTO RANDOMIZATION PROCESS		231
3-	PATIENTS LOST TO VIEW:		
	Study record lost	: 1	
	Patient died before test performed	: 6	
	Physician's refusal of randomization outcome	: 3	
	Patient's refusal of procedure	: 2	
	Procedure cancelled due to change in patient's condition	: 5	
	Patient discharged prior to procedure	: 4	
	TOTAL	: 21	21
4-	USABLE RECORDS		210

TABLE II

OUTCOME OF RANDOMIZATION PROCESS

DIAGNOSTIC GROUP OR PROTOCOL	RANDOMIZATION OUTCOME MODALITY	NO. CASES	CASES LOST TO VIEW	USABLE RECORDS
Subhepatic abscess	Ultrasound	2	0	2
	CT	1	1	0
	TOTAL	3	1	2
All other liver masses; jaundice of undiagnosed origin	Nuclear medicine	56	3	53
	Ultrasound	65	6	58
	CT	53	6	47
	TOTAL	173	15	158
Pancreatic disorders	Ultrasound	27	0	27
	CT	28	5	23
	TOTAL	55	5	50
All cases		231	21	210

patient's condition. In four other cases, the patient was discharged before the study procedure was performed; in such cases, it was not possible to ascertain whether there had been a change in the patient's condition, whether other diagnostic information had obviated the need for the diagnostic imaging procedure, or whether at that time the attending physician planned to ultimately reschedule the diagnostic imaging procedure on an ambulatory basis.

An attending physician(s) refused to have the patient undergo the diagnostic procedure to which he (she) had been randomly assigned, in three cases, all of which had been assigned to CT; the patient refused to submit to the study procedure in two instances (one had been assigned to ultrasound, the other to CT); and one study record was lost during the author's move from Montreal to Quebec City.

Consequently a total of 210 usable records were analysed and constitute the data base for the present study.

B- COMPARABILITY OF PATIENT GROUPS

The analyses presented in this chapter are based on data available for 194 of the 210 usable records. Two sets of data representing a total of 16 records were excluded, simply because the small number of cases in each of those two sets did not allow analysis; the two diagnostic groups consist of 2 patients investigated for subhepatic abscess and 14 patients investigated for jaundice of undiagnosed origin. Thus, only two diagnostic groupings remain: patients investigated for a liver mass of any origin (most of which were cases of suspected neoplastic liver disease) and patients investigated for pancreatic disorders, namely

acute or chronic pancreatitis and pancreatic cancer, pooled into one group to allow sufficiently large numbers for analysis.

Each diagnostic group is comprised of two or three protocol groups, depending on the number of imaging modalities to which subjects were randomly assigned. Data regarding comparability of protocol groups in terms of age and sex characteristics appear on Table III. When all subjects investigated for liver neoplasm are considered, there is no statistically significant difference in the mean age of subjects and the proportion of males between protocol groups. The mean age of subjects in each group varies from 60.4 years in the CT group to 64.0 years in the nuclear medicine group ($F_{2,131} = 1.06$; $p > .05$). Thus the null hypothesis (of no difference between the three groups on the average age of subjects) cannot be rejected.

Similarly when the three protocol groups within the liver neoplasm category are compared for male-female distribution, chi-square analysis ($\chi^2 = 0.27$; $p > .80$) shows no statistically significant difference between the three groups, the proportion of males varying from 19 of 44 subjects (43.2%) to 19 of 39 subjects (48.7%).

When all patients investigated for a liver mass (of an infectious, cystic or neoplastic origin) by either nuclear medicine, ultrasound or CT were compared for mean age and proportion of males to females, no statistically significant nor important difference was observed among the three groups ($F_{2,141} = 2.13$; $p > .05$; $\chi^2 = .597$; $p > .70$).

When the mean age of patients investigated for pancreatic disorders by

TABLE III

PATIENT CHARACTERISTICS ACCORDING TO PROTOCOL GROUP

PROTOCOL GROUP	N	MEAN AGE	(S.D.)	PROPORTION OF MALES
LIVER NEOPLASMS				
CT	39	60.43	(10.56)	19/39
ULTRASOUND	51	63.51	(11.97)	24/51
NUCLEAR MEDICINE	44	64.0	(13.51)	19/44
	134	$F_{2,131} = 1.06$; $p > .05$		$X_2^2 = .277$; $p > .80$
ALL LIVER MASSES				
CT	42	58.57	(12.75)	22/42
ULTRASOUND	54	63.60	(11.76)	24/54
NUCLEAR MEDICINE	48	62.94	(13.77)	23/48
	144	$F_{2,141} = 2.13$; $p > .05$		$X_2^2 = .597$; $p > .70$
PANCREATIC DISORDERS				
CT	23	62.39	(11.84)	8/23
ULTRASOUND	27	59.11	(17.42)	14/27
	50	$t_{48} = 0.76$; $p > .10$		$X_1^2 = 1.468$; $p > .20$

ultrasound was compared to that of patients those investigated by computed tomography, no statistically significant difference was found ($t_{48} = 0.76$; $p > .10$). The mean age of CT subjects was 62.39 years while that of ultrasound subjects was 59.11 years.

Similarly, chi-square analysis ($\chi^2 = 1.468$; $p > .20$) showed no statistically significant difference in the proportion of males-females between those two protocol groups, although the ultrasound group consisted of a larger proportion of males (14/27 or 51.9% of subjects) than the CT group (8/23 or 34.8% of subjects).

C- RELATIVE TECHNICAL QUALITY OF SCANS

The analysis of the various measures of diagnostic efficacy used to compare the three diagnostic imaging modalities under study and reported in this chapter, was carried out with two sets of data, for many of the study parameters. One set of data includes all cases for which a complete study record is available; the other set excludes all cases for which the imaging procedure of interest produced a "technically poor" scan which meant that the imaging specialist considered, at the time the scan was performed, that technical reasons (inadequate radioisotope uptake in the case of nuclear medicine scans, overlying bowel gas in patient's abdominal cavity in the case of ultrasound scans) did not permit acceptable visualization of the organ of interest.

All technically poor scans were considered as such (at the time the scan was performed) by the appropriate imaging specialist, and recorded as such on the written scan report.

Two reasons prompted the author to report both sets of data. First, technically poor scans cannot be expected to have an impact on a clinician's diagnostic assessment, and thus comparing the diagnostic usefulness of diagnostic technologies without excluding technically poor scans might introduce bias and create "noise" in terms of interpretation.

Secondly, it is pertinent to report data which do include technically poor scans, since a diagnostic technology's "technical efficacy" (Fineberg 1977: ability to produce a clear image allowing diagnosis to be made) is a requisite to its clinical or diagnostic efficacy as measured here.

Table IV presents the relative frequency of technically poor liver scans for each modality under study. No statistically significant difference was observed between the three technologies ($\chi^2 = 2.609$; $p > .20$). CT produced no technically poor scan (out of 42 cases), while nuclear medicine produced three (3) technically poor scans (out of 48 cases) and ultrasonography also produced three (3) technically poor scans (out of 54 cases).

In the case of scans of the pancreas, Table V indicates that there is a significant ($p < .01$) difference between the frequency of technically poor scans performed under computed tomography vs ultrasonography, the former producing no technically poor scans out of 23 cases, the latter producing eight (8) technically poor scans out of 27 cases (all of them due to overlying bowel gas).

TABLE IV

TECHNICAL QUALITY OF SCANS OF THE LIVER
AMONG STUDY PATIENTS, ACCORDING TO IMAGING MODALITY

	IMAGING MODALITY			
	Nuclear Medicine	Ultrasound	CT	Total
Technically good scans	45	51	42	138
Technically poor scans	3	3	0	6
TOTAL	48	54	42	144

$X^2 = 2.609$; DF = 2; $p > .20$

TABLE V

TECHNICAL QUALITY OF SCANS OF THE PANCREAS
AMONG STUDY PATIENTS, ACCORDING TO IMAGING MODALITY

	IMAGING MODALITY		
	Ultrasound	CT	Total
Technically good scans	19	23	42
Technically poor scans	8	0	8
TOTAL	27	23	50

$X^2 = 8.12$; DF = 1; $p < .01$

Though it may be argued that the total number of poor scans in the chi-square analysis does not quite meet the conventional requirement of an expected count of 5 per cell in such contingency tables (Armitage 1971), the exactness of the χ^2 statistic has been demonstrated (Lewontin 1965, Craddock 1970, Larntz 1978) in 2 x N tables where the expected value per cell is as small as one. In such circumstance the χ^2 statistic tends to be conservative.

D- FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD

In chapters II (literature review) and III (methods) the rationale and construct of the measure of diagnostic efficacy (i.e. the ability of a diagnostic imaging procedure's results to effect a change in the clinician's assessment of the likelihood of a diagnosis, as measured on the DAF) was described.

This section of chapter IV will present the results of various comparisons of the frequency of change in diagnostic likelihood among patients investigated for liver masses and those investigated for pancreatic disorders, according to the imaging procedure used.

1- LIVER MASSES

Tables VI-A and VI-B present data on the number of patients for whom a change in diagnostic likelihood was observed among those investigated for a liver mass (cyst, abscess, or neoplasm) by either nuclear medicine (NM), ultrasonography (US), or computed tomography (CT).

Chi-square analysis indicates that when all cases are considered, there

TABLE VI-A

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG CASES INVESTIGATED FOR A LIVER MASS,
ACCORDING TO PROTOCOL GROUP
(ALL CASES)

	PROTOCOL GROUP			
	Nuclear Medicine	Ultrasound	CT	Total
Change in diagnostic likelihood	26	24	26	76
No change in diagnostic likelihood	22	30	16	68
TOTAL	48	54	42	144

$X^2 = 2.25$; DF = 2; $p > .10$

TABLE VI-B

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG CASES INVESTIGATED FOR A LIVER MASS,
ACCORDING TO PROTOCOL GROUP
(TECHNICALLY POOR SCANS EXCLUDED)

	PROTOCOL GROUP			
	Nuclear Medicine	Ultrasound	CT	Total
Change in diagnostic likelihood	26	24	26	76
No change in diagnostic likelihood	19	27	16	62
TOTAL	45	51	42	138

$X^2 = 2.25$; DF = 2; $p > .10$

is no significant difference in the relative frequency of change in diagnostic likelihood when the three imaging modalities are compared ($\chi^2 = 2.94$; $p > .10$). Therefore, the null hypothesis of no difference in frequency of change in diagnostic likelihood among the three imaging techniques cannot be rejected. Similar analysis excluding the technically poor scans (which cannot be expected to influence diagnostic likelihood) yielded similar results ($\chi^2 = 2.25$; $p > .10$).

Since case-mix and heterogeneity of probable diagnoses could dilute the effect of one or more of the diagnostic technologies under study and contribute to the results of the last two tables, the same analysis was repeated with only those patients investigated for (primary or secondary) neoplasm of the liver. This provides a more homogeneous group since 124 of the 133 cases in this group were being investigated for metastases to the liver. Again, no significant difference in the frequency of change in diagnostic likelihood can be observed among the three groups of patients, as illustrated in Tables VII-A ($\chi^2 = 2.71$; $p > .10$) and VII-B ($\chi^2 = 1.88$; $p > .10$).

Information theory suggests that it is possible that the initial likelihood level could influence the probability of a change in diagnostic likelihood subsequent to a scan (of whichever nature); thus, a potential difference (or lack of it) in the frequency of change in diagnostic likelihood could be confounded by differences in the distribution of cases according to the initial diagnostic likelihood level.

This possibility was tested by the hypothesis that there is no significant difference in frequency of change in diagnostic likelihood among

TABLE VII-A

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG CASES INVESTIGATED FOR LIVER NEOPLASM
ACCORDING TO PROTOCOL GROUP
(ALL CASES)

	PROTOCOL GROUP			
	Nuclear Medicine	Ultrasound	CT	Total
Change in diagnostic likelihood	23	22	24	69
No change in diagnostic likelihood	21	28	15	64
TOTAL	44	50	39	133

$X^2 = 2.71$; DF = 2; $p > .10$

TABLE VII-B

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG CASES INVESTIGATED FOR LIVER NEOPLASM
ACCORDING TO PROTOCOL GROUP
(EXCLUDING TECHNICALLY POOR SCANS)

	PROTOCOL GROUP			
	Nuclear Medicine	Ultrasound	CT	Total
Change in diagnostic likelihood	23	22	24	69
No change in diagnostic likelihood	19	25	15	59
TOTAL	42	47	39	128

$X^2 = 1.88$; DF = 2; $p > .10$

TABLE VIII-A

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG PATIENTS INVESTIGATED FOR LIVER MASSES
ACCORDING TO INITIAL DIAGNOSTIC LIKELIHOOD LEVEL
(ALL CASES)

	INITIAL LIKELIHOOD LEVEL				
	(1) Very Unlikely	(2) Unlikely	(3) Likely	(4) Very Likely	Total
Change in diagnostic likelihood	1	36	31	8	76
No change in diagnostic likelihood	41	13	7	7	68
TOTAL	42	49	38	15	144

$X^2 = 63.88$; DF = 3; $p < .001$

TABLE VIII-B

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG PATIENTS INVESTIGATED FOR LIVER MASSES
ACCORDING TO INITIAL DIAGNOSTIC LIKELIHOOD LEVEL
(EXCLUDING TECHNICALLY POOR SCANS)

	INITIAL LIKELIHOOD LEVEL				
	(1) Very Unlikely	(2) Unlikely	(3) Likely	(4) Very Likely	Total
Change in diagnostic likelihood	1	36	31	8	76
No change in diagnostic likelihood	40	9	6	7	62
TOTAL	41	45	37	15	138

$X^2 = 69.07$; DF = 3; $p < .001$

groups of patients investigated for a liver mass when these patients were grouped according to the initial pre-scan diagnostic likelihood level recorded.

According to the results of the chi-square analysis shown in Table VIII-A, a large and significant difference in the frequency of change in diagnostic likelihood exists ($\chi^2 = 63.88$; $p < .001$) among patients investigated for liver mass grouped according to initial likelihood level. Indeed, of the 42 cases for whom the initial diagnosis was considered very unlikely, only one recorded a change in diagnostic likelihood subsequent to the study scan, while the frequency of change at other likelihood levels was 36/49, 31/38, and 8/15 respectively. Similar results were obtained when technically poor scans were excluded from the analysis (Table VIII-B). Thus, in cases investigated for liver mass, a clinician's initial diagnostic likelihood assessment is less likely to change subsequent to a nuclear medicine, ultrasound or CT scan, when the diagnosis under investigation is considered very unlikely, prior to the scan. Hence, if the proportion of cases in the very unlikely category is not the same among the diagnostic imaging groups, it is possible that a bias exists in the data presented in Tables VI-A & B and VII-A & B. A contingency table was constructed using the frequency of very unlikely initial diagnosis vs other likelihood levels as response variable, and the imaging modality as stimulus variable.

Results of the analysis presented in Table IX indicate there is no significant difference in the frequency of an initially very unlikely diagnosis (of liver mass) among the patients investigated by any one of the three diagnostic imaging modalities under study, although the

TABLE IX

FREQUENCY OF VERY UNLIKELY INITIAL DIAGNOSIS
AMONG PATIENTS INVESTIGATED FOR A LIVER MASS
ACCORDING TO PROTOCOL GROUP
(ALL CASES)

INITIAL DIAGNOSTIC LIKELIHOOD LEVEL	DIAGNOSTIC IMAGING MODALITY			
	Nuclear Medicine	Ultrasound	CT	Total
Very unlikely	15	17	10	42
All other likelihood levels	33	37	32	102
TOTAL	48	54	42	144

$X^2 = 0.298$; DF = 2; $p > .80$

proportion of such cases is slightly smaller in the CT group. Hence, there is no reason to assume that the initial diagnostic likelihood level was an important confounding variable in the analysis reported on Tables VI and VII. In summary, there is no significant difference between nuclear medicine, ultrasonography or computed tomography as regards the relative number of patients for whom a change in the physician's diagnostic likelihood estimate was observed, subsequent to the scan report.

2- PANCREATIC DISORDERS

As in the preceeding section, patients investigated for pancreatic disorders (pancreatitis or pancreatic neoplasm) by either of the imaging modalities under study, ultrasonography or CT were compared as to the relative frequency of cases where a change in the clinician's diagnostic likelihood assessment occurred subsequent to information provided by the study scan. Again, analysis of this efficacy parameter was done with all cases, then repeated with data excluding technically poor scans.

Results appearing in Table X-A indicate that when all cases are considered, there is a considerable and significant ($p < .01$) difference between the two groups of patients in the relative frequency of change in diagnostic likelihood; cases investigated by computed tomography showed a much greater post-scan frequency of change in diagnostic likelihood (16/23 cases) than did those investigated by ultrasonography (only 8/27 cases). It should be noted, however, that seven of 19 ultrasound scans which were not followed by a change in diagnostic likelihood were classified as "technically poor scans", and consequently, not expected to influence diagnostic assessment since they provided no pertinent diagnostic information.

TABLE X-A

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG CASES INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO PROTOCOL GROUP
(ALL CASES)

	PROTOCOL GROUP		
	Ultrasound	CT	Total
Change in diagnostic likelihood	8	16	24
No change in diagnostic likelihood	19	7	26
TOTAL	27	23	50

$X^2 = 7.94$; DF = 1; $p < .01$

TABLE X-B

FREQUENCY OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG CASES INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO PROTOCOL GROUP
(EXLUDING TECHNICALLY POOR SCANS)

	PROTOCOL GROUP		
	Ultrasound	CT	Total
Change in diagnostic likelihood	7	16	23
No change in diagnostic likelihood	12	7	19
TOTAL	19	23	42

$X^2 = 4.48$; DF = 1; $p < .05$

The chi-square analysis was thus repeated after all "technically poor scans" were removed from the data. According to the results on Table X-B, cases investigated for pancreatic disorders by computed tomography incurred a significantly greater relative frequency of post-scan change in diagnostic likelihood than did those investigated by ultrasonography, even after excluding technically poor scans.

It can thus be concluded, that CT contributes to a change in the clinician's diagnostic likelihood assessment in a significantly greater proportion of patients investigated for pancreatic disorders (pancreatitis and pancreatic neoplasm), than ultrasonography in similar cases, whether or not technically poor scans are excluded.

E- MEAN DIAGNOSTIC EFFICACY SCORE

The second measure of diagnostic efficacy for which the competing imaging modalities under study were compared, was the mean diagnostic efficacy score of each modality in each patient group, i.e. liver masses and pancreatic disorders.

As described in detail in chapter III (methods) the diagnostic efficacy score of any scan in a given case is a weighted value of the post-scan change in diagnostic likelihood, or lack of it, designed to reflect numerically the fact that diagnostic likelihood changes of similar magnitude on the 4-point likelihood scale may have a very different decision-making value (importance) for the clinician at different points on the scale.

1- LIVER MASSES

When the diagnostic efficacy scores of each of the three imaging techniques for all patients investigated for a liver mass were tabulated and their respective mean values calculated, nuclear medicine had a mean score of 1.222 (45 cases), ultrasound obtained a mean score of 0.804 (51 cases), and computed tomography's mean diagnostic efficacy score was 1.071 (42 cases).

One-way analysis of variance (Table XI) indicates that one cannot reject the null hypothesis of no significant difference among the three groups as to their respective mean diagnostic efficacy score ($F_{2,135} = .18$; $p > .05$).

Given the fact that previous analysis (Tables VI-A & B) showed no significant between-group difference in the frequency of change in diagnostic likelihood among patients investigated for a liver mass and that cases where no change in likelihood occurred, were included in the analysis presented on Table XI, it is not surprising that no significant difference in mean diagnostic efficacy score was observed, as reported on Table XI.

However, it is still possible, despite the finding of no significant difference in frequency of change in likelihood among NM, US and CT, that one of these techniques contributed to a significantly greater or smaller change when change in diagnostic likelihood occurred, a phenomenon which could not be identified from the analysis reported in Table XI.

Consequently, analysis of variance was repeated with the same data but

TABLE XI

DIAGNOSTIC EFFICACY SCORES AMONG PATIENTS INVESTIGATED FOR LIVER MASS ACCORDING TO PROTOCOL GROUP (EXCLUDING TECHNICALLY POOR SCANS)

BEFORE-AFTER READING on Diagnostic Likelihood Scale	EFFICACY VALUE (points)	PROTOCOL GROUP					
		NUCLEAR MEDICINE		ULTRASOUND		CT	
		Frequency	Score	Frequency	Score	Frequency	Score
1-1	0	15	0	15	0	10	0
1-2	2	-	-	-	-	-	-
1-3	3	-	-	1	3	-	-
1-4	5	-	-	-	-	-	-
2-1	1	10	10	10	10	12	12
2-2	0	2	0	6	0	1	0
2-3	1	-	-	1	1	2	2
2-4	3	-	-	-	-	1	3
3-1	3	9	27	4	12	5	15
3-4	1	3	3	4	4	2	2
3-2	0	1	0	2	0	3	0
3-4	2	1	2	2	4	1	2
4-1	5	2	10	1	5	-	-
4-2	3	1	3	-	-	3	9
4-3	2	-	-	1	2	-	-
4-4	0	1	0	4	0	2	0
TOTAL		45	55	51	41	42	45
MEAN SCORE PER CASE		1.222		0.804		1.071	
STANDARD DEVIATION		1.428		1.132		1.135	
VARIANCE		2.04		1.28		1.29	

$F = 0.18$; $V_1 = 2$ DF; $V_2 = 135$ DF; $p > .05$

excluding cases where the diagnostic efficacy score was zero (no change in diagnostic likelihood). The results appearing on Table XII indicate that when these cases alone are considered, there is still no significant difference among the mean diagnostic efficacy scores of nuclear medicine (mean score 2.115 for 26 cases), ultrasound (mean score 1.708 for 24 cases) and CT (mean of 1.731 for 26 cases) ($F_{2,73} = .066$; $p > .05$).

In concluding (from Tables VI, XI and XII) it can be said that when nuclear medicine, ultrasound and CT are used in the investigation of similar patients for liver masses, there is no significant difference in the proportion of such cases in which the clinician's diagnostic likelihood assessment will be influenced by the results of the scan, nor does there appear to be any difference among them as to the importance of a change in diagnostic likelihood when it does occur.

2- PANCREATIC DISORDERS

Similarly, when the mean diagnostic efficacy scores of ultrasound and CT used in the investigation of pancreatic disorders are compared for all cases (except technically poor scans), no significant difference was observed between the two means ($t_{40} = 0.68$; $p > 50$; Table XIII).

As in the case of liver masses, the present analysis was repeated with the same data but excluding cases with a diagnostic efficacy score of zero (no change in diagnostic likelihood). As reported in Table XIV, no significant difference was observed between the mean diagnostic efficacy score of ultrasound and that of CT among those cases where a change in likelihood occurred ($t_{21} = 0.58$; $p > .10$).

TABLE XII

DIAGNOSTIC EFFICACY SCORES ACCORDING TO PROTOCOL GROUP AMONG PATIENTS INVESTIGATED FOR A LIVER MASS AND WHERE CHANGE IN DIAGNOSTIC LIKELIHOOD OCCURRED SUBSEQUENT TO IMAGING PROCEDURE

CHANGE OBSERVED on Diagnostic Likelihood Scale	EFFICACY VALUE (Points)	PROTOCOL GROUP					
		NUCLEAR MEDICINE		ULTRASOUND		CT	
		Frequency	Score	Frequency	Score	Frequency	Score
1-2	2	-	-	-	-	-	-
1-3	3	-	-	1	3	-	-
1-4	5	-	-	-	-	-	-
2-1	1	10	10	10	10	12	12
2-3	1	-	-	1	1	2	2
2-4	3	-	-	-	-	1	3
3-1	3	9	27	4	12	5	15
3-2	1	3	3	4	4	2	2
3-4	2	1	2	2	4	1	2
4-1	5	2	10	1	5	-	-
4-2	3	1	3	-	-	3	9
4-3	2	-	-	1	2	-	-
TOTAL		26	55	24	41	26	45
MEAN SCORE PER CASE		2.115		1.708		1.731	
STANDARD DEVIATION		1.275		1.080		0.96	
VARIANCE		1.626		1.172		0.92	

$F_{2,73} = .066$; $p > .05$

TABLE XIII

DIAGNOSTIC EFFICACY SCORES AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS ACCORDING TO PROTOCOL GROUP (EXCLUDING TECHNICALLY POOR SCANS)

BEFORE-AFTER READING on Diagnostic Likelihood Scale	EFFICACY VALUE (Points)	PROTOCOL GROUP			
		ULTRASOUND		CT	
		Frequency	Score	Frequency	Score
1-1	0	6	0	3	0
1-2	2	-	-	1	2
1-3	3	-	-	1	3
1-4	5	-	-	-	-
2-1	1	1	1	3	3
2-2	0	-	-	1	0
2-3	1	-	-	1	1
2-4	3	-	-	1	3
3-1	3	1	3	5	15
3-2	1	-	-	1	1
3-3	0	2	0	-	-
3-4	2	2	4	2	4
4-1	5	-	-	-	-
4-2	3	2	6	1	3
4-3	2	1	2	-	-
4-4	0	4	0	3	0
TOTAL		19	16	23	35
MEAN SCORE PER CASE		0.842		1.522	
STANDARD DEVIATION		1.214		1.275	
VARIANCE		1.474		1.625	

$t_{40} = 0.68$; $p > 0.50$

TABLE XIV

DIAGNOSTIC EFFICACY SCORES ACCORDING TO PROTOCOL GROUP AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS WHERE CHANGE IN DIAGNOSTIC LIKELIHOOD OCCURRED SUBSEQUENT TO IMAGING PROCEDURE

CHANGE OBSERVED on Diagnostic Likelihood Scale	EFFICACY VALUE (Points)	PROTOCOL GROUP			
		ULTRASOUND		CT	
		Frequency	Score	Frequency	Score
1-2	2	-	-	1	2
1-3	3	-	-	1	3
1-4	5	-	-	-	-
2-1	1	1	1	3	3
2-3	1	-	-	1	1
2-4	3	-	-	1	3
3-1	3	1	3	5	15
3-2	1	-	-	1	1
3-4	2	2	4	2	4
4-1	5	-	-	-	-
4-2	3	2	6	1	3
4-3	2	1	2	-	-
TOTAL		7	16	16	35
MEAN SCORE PER CASE		2.286		2.188	
STANDARD DEVIATION		0.756		0.911	
VARIANCE		0.571		0.829	

$t_{21} = 0.58$; $p > .10$

In concluding from Tables X, XIII and XIV, it can be said that although CT seems to influence a clinician's diagnostic likelihood assessment in a significantly greater proportion of patients investigated for pancreatic disorders than does ultrasound, there is no significant difference between the two technologies regarding the importance (mean value) of changes in diagnostic likelihood when they do occur as a result of the scan.

F- IMPROVEMENT IN DIAGNOSTIC UNDERSTANDING

Comparative analyses reported in the previous sections have concentrated on the influence of the diagnostic technologies under study on the clinicians' estimate of the likelihood of a suspected diagnosis. The technologies were compared on the basis of the frequency with which each of them contributed to a change in the clinician's diagnostic likelihood estimate as well as the magnitude of the change when such change occurred.

However, changes in diagnostic likelihood can reflect either an improvement or a deterioration in the clinician's understanding of the patient's condition. Thus, it is possible that a diagnostic procedure which frequently alters diagnostic likelihood estimates only contributes to confuse the clinician's diagnostic understanding and vice-versa.

Given the construct of the diagnostic likelihood scale, it is fair to assume that only those diagnostic imaging procedures that result in changing likelihood estimates toward either of the two extreme points on the scale, i.e. very unlikely and very likely, contribute to changing the clinician's diagnostic understanding for the better, allowing him to either confirm or rule out a suspected diagnosis.

Furthermore, since the two extreme points on the four-point diagnostic likelihood scale indicate either a very low (less than 25%) or very high (75% or more) probability of presence of the suspected pathology, it is reasonable to assume that patient management decisions made at those levels are more determinate. Consequently, the results of a diagnostic imaging procedure which would shift the initial diagnostic likelihood estimate toward either of these two extreme points on the scale, would also have a more determinate effect on subsequent patient management.

In order to appreciate the improvement in diagnostic understanding provided by each technology on a comparative basis, an analysis was performed in two parts.

In the first part of the analysis, comparison of pre-scan versus post-scan frequency of extreme likelihood levels in each protocol group and between protocol groups was used to provide an overall picture of the contribution of each diagnostic modality towards increasing the proportion of diagnostically definite cases within its patient group.

The second part of the analysis included only those cases where a change in the physician's diagnostic likelihood estimate occurred as a result of the scan; the nature of the change in likelihood was identified as "improvement" whenever that change from either of the two central points on the scale (likely or unlikely) to either of the extreme points on the scale (very likely or very unlikely). Changes from either of the extreme points on the scale (very likely or very unlikely) to either of the two central points on the scale (likely or unlikely) were considered as

"losses" since they illustrated an actual reduction in diagnostic certainty; finally, those changes from one central point on the scale to the other, as well as those changes from one extreme point on the scale to the other were considered as "neutral" since neither showed an improvement nor a loss in diagnostic certainty. It is important to emphasize that a change from very unlikely to very likely is an important change with a large effect on subsequent decisions regarding patient management. However, such a change in diagnostic understanding cannot a priori be qualified as an "improvement" in the physician's understanding of the patient's condition, since the diagnostic certainty has not changed, and since at the time the scan findings were reported to the clinician, there was no proof of the correctness of the finding. Thus, changes from one extreme point to the other, or from one central point to the other on the diagnostic likelihood scale were considered "neutral". Because of the small number of cases, "neutral" and "loss" cases were pooled and reported as "no improvement" in diagnostic understanding.

Both phases of this analysis are reported for the liver mass group and the pancreatic group on Tables XV to XXII.

1 - LIVER MASSES

When the proportion of cases (investigated for liver mass) with extreme diagnostic likelihoods before the scan is compared to the proportion of similar cases after the scan, an important and significant increase is observed subsequent to the scan for each of the three imaging procedures under study. All technically poor and non-diagnostic scans were excluded from the analysis since neither could be expected to shift

diagnostic likelihood estimates toward or away from the extreme points on the likelihood scale.

As reported in Table XV, the proportion of cases with an extreme diagnostic likelihood level doubled from 19/45 to 38/45 in the group of patients submitted to nuclear medicine, and this increase was found to be significant ($\chi^2 = 17.28$; $p < .001$); it increased significantly from 22/50 to 35/50 after the scan among those patients investigated by ultrasound ($\chi^2 = 8.20$; $p < .01$); and it doubled from 15/42 to 31/42 after the scan, among patients investigated by computed body tomography, again the increase was found to be statistically significant ($\chi^2 = 12.30$; $p < .001$).

It is thus apparent that when cases investigated for a liver mass were considered, all three diagnostic imaging modalities under study contributed significantly and substantially toward increasing the proportion of diagnostically definite cases (either by confirming or ruling out a diagnostic possiblity).

When the three imaging modalities were subsequently compared (using log-linear analysis) for change in the proportion of diagnostically definite cases based on a between-group comparison of the pre-scan vs post-scan distribution of likelihood levels, according to the data presented in Table XVI, no significant between-group difference was observed. The G^2 statistic obtained ($G^2 = 22.8$; $p > .05$) confirmed the null hypothesis of total independence between technology groups in their respective before versus after change in the proportion of diagnostically definite cases.

The second phase of the analysis regarding improvement in diagnostic

TABLE XV

FREQUENCY OF EXTREME LIKELIHOOD LEVELS
BEFORE AND AFTER SCAN RESULTS
AMONG PATIENTS INVESTIGATED FOR LIVER MASS
(EXCLUDING TECHNICALLY POOR AND NON-DIAGNOSTIC SCANS)

NUCLEAR MEDICINE

	Before	After
Extreme levels	19	38
Others	26	7
TOTAL	45	45

$X^2 = 17.28$; DF = 1; $p < .001$

ULTRASOUND

	Before	After
Extreme levels	22	35
Others	28	15
TOTAL	50	50

$X^2 = 8.20$; DF = 1; $p < .01$

CT

	Before	After
Extreme levels	15	31
Others	27	11
TOTAL	42	42

$X^2 = 12.30$; DF = 1; $p < .001$

TABLE XVI

DISTRIBUTION OF PRIOR AND POSTERIOR DIAGNOSTIC LIKELIHOODS AMONG PATIENTS INVESTIGATED FOR LIVER MASS ACCORDING TO IMAGING MODALITY

NUCLEAR MEDICINE

Posterior Diagnostic Likelihood	Prior Diagnostic Likelihood 1	2	3	4
1	15	10	9	2
2	-	2	3	1
3	-	-	1	-
4	-	-	1	1

ULTRASOUND

Posterior Diagnostic Likelihood	Prior Diagnostic Likelihood 1	2	3	4
1	15	10	4	1
2	-	6	4	-
3	1	1	2	1
4	-	-	1	4

CT

Posterior Diagnostic Likelihood	Prior Diagnostic Likelihood 1	2	3	4
1	10	12	5	-
2	-	1	2	3
3	-	2	3	-
4	-	1	1	2

understanding focused specifically on only those cases where a change in diagnostic likelihood was recorded. The purpose of this second phase was to identify on a case-by-case basis whether a given diagnostic imaging modality contributed or not to an improvement of diagnostic understanding, when it was shown to have contributed to a change in the physician's diagnostic likelihood estimates. The definition of the variable "improvement in diagnostic understanding" has been outlined earlier in this section.

According to the results appearing in Table XVII, there was no statistically significant nor numerically important difference between the nuclear medicine, ultrasound and CT groups as regards the proportion of cases where diagnostic understanding was improved, among those cases where a change in diagnostic likelihood was recorded (excluding technically poor and non-diagnostic scans). An improvement in diagnostic understanding was recorded in 21 of 27 cases in the NM group, in 15 of 23 cases in the US group, and in 19 of 26 cases in the CT group ($\chi^2 = 0.987$; $p > .50$).

Hence, it can be concluded that no significant difference exists between nuclear medicine, ultrasound and CT in their respective capacity to increase the overall proportion of diagnostically definite cases in a given group of patients investigated for a liver mass (all three have an important effect in that respect); and that no significant difference was found between the three imaging modalities in the proportion of cases where diagnostic understanding was actually improved when the modality lead to a change in diagnostic likelihood.

TABLE XVII

NATURE OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG PATIENTS INVESTIGATED FOR LIVER MASS
ACCORDING TO IMAGING MODALITY*

Nature of Change in Diagnostic Likelihood	IMAGING MODALITY NM	US	CT	Total
Improvement	21	15	19	55
No improvement	6	8	7	21
TOTAL	27	23	26	76

$X^2 = 0.987$; DF = 2; $p > .50$

* Excluding technically poor and non-diagnostic scans.

2- PANCREATIC DISORDERS

Similarly, in the first phase of analysis of this efficacy indicator among patients investigated for pancreatic disorders, the pre-scan frequency of extreme diagnostic likelihood levels was compared to the post-scan frequency in each of the two imaging groups (excluding technically poor and non-diagnostic scans).

Among the 18 patients in the ultrasound group for whom the scan produced a clinical finding, the proportion of cases with extreme diagnostic likelihood increased from 12/18 before the scan to 14/18 after the ultrasound scan. This difference however was not found to be statistically significant ($\chi^2 = 0.552$; $p > .30$; Table XVIII).

When the post-scan frequency of extreme diagnostic likelihood levels was compared to the pre-scan frequency among the 20 patients for whom a CT scan of the pancreas produced a clinical finding (Table XIX), an important and statistically significant difference ($p < .02$) was observed (7/20 cases before the CT scan compared to 15/20 cases after the scan).

These analyses suggest that only CT (and not ultrasound) had an important and statistically significant effect in increasing the overall proportion of diagnostically definite cases.

As described previously, the second phase of the analysis focused on the proportion of "better" diagnoses among cases where a change in diagnostic likelihood had been observed subsequent to the scan, based on the aforementioned definition of improvement in diagnostic understanding, and on the pre-scan vs post-scan distributions of likelihood levels appearing

TABLE XVIII

PRE-SCAN VS POST-SCAN FREQUENCY OF EXTREME DIAGNOSTIC LIKELIHOOD LEVELS AMONG ULTRASOUND SCANS OF THE PANCREAS*

	BEFORE SCAN	AFTER SCAN
Frequency of extreme diagnostic likelihood	12	14
Frequency of other diagnostic likelihood	6	4
TOTAL	18	18

$X^2 = 0.552$; DF = 1; $p > .30$

* Excluding technically poor and non-diagnostic scans.

TABLE XIX

PRE-SCAN VS POST-SCAN FREQUENCY OF EXTREME DIAGNOSTIC LIKELIHOOD LEVELS AMONG CT SCANS OF THE PANCREAS*

	BEFORE SCAN	AFTER SCAN
Frequency of extreme diagnostic likelihood	7	15
Frequency of other diagnostic likelihood	13	5
TOTAL	20	20

$X^2 = 6.466$; DF = 1; $p < .02$

*Excluding technically poor and non-diagnostic scans

TABLE XX

DISTRIBUTION OF PRIOR AND POSTERIOR DIAGNOSTIC LIKELIHOODS AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS ACCORDING TO IMAGING MODALITY

CT

POSTERIOR DIAGNOSTIC LIKELIHOOD	PRIOR DIAGNOSTIC LIKELIHOOD			
	1	2	3	4
1	3	2	5	-
2	1	1	1	-
3	1	1	-	-
4	-	1	2	2

US

POSTERIOR DIAGNOSTIC LIKELIHOOD	PRIOR DIAGNOSTIC LIKELIHOOD			
	2	2	3	4
1	6	1	1	-
2	-	-	-	1
3	-	-	3	-
4	-	-	2	4

in Table XX.

Table XXI shows that the frequency of cases where diagnostic understanding was improved as a result of the scan was greater in the CT group, (10 of 20 cases), than in the ultrasound group only 4 of 18 cases (χ^2 = 3.137; p slightly larger than .05).

It can thus be concluded that when applied in the investigation of pancreatic disorders, CT significantly increases the proportion of diagnostically definite cases while US does not, and that in cases where CT contributes to changing physicians' diagnostic likelihood estimates, such changes are more likely to lead to improved diagnostic understanding.

G- FREQUENCY OF SUPPLEMENTARY DIAGNOSTIC PROCEDURE

This section deals with the comparison of the competing technologies of interest with respect to the frequency of supplementary diagnostic (imaging) procedures performed for similar indications, subsequent to reception of the initial scan's report. Only those supplementary diagnostic procedures pertaining specifically to the diagnostic indication of the original scan were included in the analysis.

1- LIVER MASSES

Table XXII-A shows that for all cases only 10/48 patients in the nuclear medicine group underwent further diagnostic procedure(s) as opposed to 37/54 in the ultrasound group and 23/42 in the CT group (χ^2 = 24.07; DF = 2; p < .001).

Since technically poor scans by definition do not provide useful diagnostic information and are thus more likely to be followed by supplementary diagnostic procedure(s), thus introducing a possible bias, and since such scans are not evenly distributed among the three technologies under study, the analysis was repeated with data excluding technically poor scans.

Results appearing on Table XXII-B indicate again that the relative number of patients undergoing supplementary testing is smaller in the NM group than in the US and CT groups ($\chi^2 = 28.46$; DF = 2, $p < .001$).

Study subjects were randomized to one of the three technologies. Thus, for the liver mass group, there was a one-in-three chance that the study scan was not that which was originally requested by the responsible physician. The fact that the scan performed may not have been the scan requested may have influenced the physician's decision to order further tests (i.e. to order subsequently the initially requested scan). Data regarding the frequency of supplementary testing among patients investigated for a liver mass, according to nature of the initially requested scan and protocol group are shown in Table XXIII.

Analysis of this effect is reported in Table XXIV and shows that the frequency of supplementary testing is importantly and significantly smaller among scans which were the same as initially requested (11/51), than among scans which were different from that initially requested (53/86) ($\chi^2 = 20.62$; $p < .001$).

However, of the 137 initial requests for scans, 126 were requests for NM,

TABLE XXI

NATURE OF CHANGE IN DIAGNOSTIC LIKELIHOOD
AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO IMAGING MODALITY*

NATURE OF CHANGE IN DIAGNOSTIC LIKELIHOOD	IMAGING MODALITY US	CT	TOTAL
Improvement	4	10	14
No improvement	14	10	24
TOTAL	18	20	38

$X^2 = 3.137$; DF = 1; $.10 > p > .05$

TABLE XXII-A

FREQUENCY OF SUPPLEMENTARY TESTING
AMONG PATIENTS INVESTIGATED FOR LIVER MASS
ACCORDING TO PROTOCOL GROUP (ALL CASES)

	PROTOCOL GROUP			
	NUCLEAR MEDICINE	ULTRASOUND	CT	TOTAL
With supplementary testing	10	37	23	70
Without supplementary testing	38	17	19	74
TOTAL	48	54	42	144

$X^2 = 24.07$; DF = 2; $p < .001$

TABLE XXII-B

FREQUENCY OF SUPPLEMENTARY TESTING
AMONG PATIENTS INVESTIGATED FOR LIVER MASS
ACCORDING TO PROTOCOL GROUP (EXCLUDING TECHNICALLY POOR SCANS)

	PROTOCOL GROUP			
	NUCLEAR MEDICINE	ULTRASOUND	CT	TOTAL
With supplementary testing	7	35	23	65
Without supplementary testing	38	16	19	73
TOTAL	45	51	42	138

$X^2 = 28.46$; DF = 2; $p < .001$

TABLE XXIII

FREQUENCY OF SUPPLEMENTARY TESTING AMONG PATIENTS INVESTIGATED FOR A LIVER MASS, ACCORDING TO NATURE OF INITIALLY REQUESTED SCAN AND PROTOCOL GROUP (EXCLUDING TECHNICALLY POOR AND NON-DIAGNOSTIC SCANS)

	PROTOCOL GROUP						
	NM		US		CT		
Nature of initially requested scan	Patients with supplem. testing	Patients without supplem. testing	Patients with supplem. testing	Patients without supplem. testing	Patients with supplem. testing	Patients without supplem. testing	Total
NM	7	38	30	14	20	17	126
US	0	0	4	2	3	2	11
CT	0	0	0	0	0	0	0
TOTAL	7	38	34	16	23	19	137

TABLE XXIV

FREQUENCY OF SUPPLEMENTARY TESTING AMONG PATIENTS INVESTIGATED FOR A LIVER MASS ACCORDING TO CONCORDANCE OF SCAN PERFORMED TO INITIALLY REQUESTED SCAN (EXCLUDING TECHNICALLY POOR AND NON-DIAGNOSTIC SCANS)

	CONCORDANCE TO INITIAL SCAN REQUEST		
	YES	NO	TOTAL
With supplementary testing	11	53	64
Without supplementary testing	40	33	73
TOTAL	51	86	137

$X^2 = 20.62$; DF = 1; $p < .001$

11 for US and none for CT. In addition, of the 11 initial requests for ultrasound, none happened to be randomized to NM. It is, thus, extremely difficult to separate the effect of the scan on supplementary testing (management efficacy) from the effect due to randomization to an other-than-requested scan.

Table XXV shows that for US, there was no difference in the frequency of requests for supplementary tests between the cases where the test scan was the same as the requested scan (4/6) compared to the cases where the test scan was other than that initially requested (30/44) (χ^2 = .006; $p > .50$).

In conclusion, the frequency of patients submitted to supplementary diagnostic procedures was significantly less among patients investigated by NM. Because most of the initial scan requests were for NM, and because no initial scan requests were for CT, it was not possible to distinguish the management efficacy of the technologies from an effect due to the study itself, in the overall analysis.

However, in the US group, such a study effect was not apparent. It thus appears likely that the management efficacy of NM was substantially better than that of US and CT.

2- PANCREATIC DISORDERS

Results reported on Table XXVI-A indicate that when all cases investigated for pancreatic disorders are considered, there is no significant difference between the ultrasound group as to the proportion of patients who had supplementary diagnostic procedures subsequent to the study

TABLE XXV

FREQUENCY OF SUPPLEMENTARY TESTING SUBSEQUENT TO AN ULTRASOUND SCAN AMOUNG PATIENTS INVESTIGATED FOR LIVER MASS ACCORDING TO CONCORDANCE OF SCAN TO INITIAL IMAGING REQUEST (EXCLUDING TECHNICALLY POOR AND NON-DIAGNOSTIC SCANS)

	CONCORDANCE TO INITIAL IMAGING REQUEST		
	YES	NO	TOTAL
Patients with supplementary testing	4	30	34
Patients without supplementary testing	2	14	16
TOTAL	6	44	50

$X^2 = .006$; DF = 1; $p > .50$

TABLE XXVI-A

FREQUENCY OF SUPPLEMENTARY TESTING
AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO PROTOCOL GROUP
(ALL CASES)

	PROTOCOL GROUP		
	ULTRASOUND	CT	TOTAL
With supplementary testing	13	11	24
Without supplementary testing	14	12	26
TOTAL	27	23	50

$X^2 = .0001$; 1 DF; $p > .70$

TABLE XXVI-B

FREQUENCY OF SUPPLEMENTARY TESTING
AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO PROTOCOL GROUP
(EXCLUDING TECHNICALLY POOR AND NON-DIAGNOSTIC SCANS)

	PROTOCOL GROUP		
	ULTRASOUND	CT	TOTAL
With supplementary testing	7	9	16
Without supplementary testing	11	11	22
TOTAL	18	20	38

$X^2 = .145$; 1 DF; $p > .70$

scan (χ^2 = .0001; DF = 1; p > .70). Even when technically poor and non-diagnostic scans are excluded from the data, the null hypothesis cannot be rejected (χ^2 = 0.145; DF = 1; p > .70).

Also as in the previous section, the effect of concordance of scan to initial scan request was analyzed. As shown in Table XXVII, 8 of 21 patients whose scan was similar to that initially requested, underwent additional testing, while 8 of 17 patients whose scan differed from the one initially requested by the responsible physician, were submitted to supplementary testing. The difference is small and not statistically significant (χ^2 = 0.309; DF = 1; p > .50). Hence, concordance cannot be considered to have an effect on the relative frequency of supplementary testing and consequently there seems to be no difference between US and CT on this indicator of management efficacy.

H- TYPE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE

As another measure of relative management efficacy, the diagnostic technologies of interest were compared on the basis of the type of the first supplementary post-scan diagnostic procedure. It was of particular interest to find out whether, in cases where supplementary testing was done, physicians would first use one of the other competing technologies included in the present study, or rather select a complementary test. For the purpose of analysis, first supplementary procedures were grouped into the following categories:

- repeat procedure
- other study procedure: NM, US or CT for liver masses: US or CT for pancreatic disorders
- non-study imaging procedure: any diagnostic imaging procedure other than study scans: abdominal or gastro-intestinal series, intravenous cholangiogram, ERCP,

TABLE XXVII

FREQUENCY OF SUPPLEMENTARY TESTING
AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO CONCORDANCE TO INITIALLY REQUESTED SCAN

	CONCORDANCE TO INITIAL SCAN REQUEST		
	YES	NO	TOTAL
With supplementary testing	8	8	16
Without supplementary testing	13	9	22
TOTAL	21	17	38

$X^2 = 0.309$; DF = 1; $p > .50$

transhepatic cholongiography, and arteriography
- diagnostic surgery
- biopsy performed independently of any other surgery

The analysis was limited only to the first post-scan supplementary diagnostic procedure since the nature of any subsequent test may just as likely be influenced by the diagnostic efficacy of the previous post-scan supplementary test as by that of the initial scan itself. It was not possible nor pertinent, given the purpose of the present study and the nature of the data collected, to attempt to isolate the effect of subsequent supplementary tests.

1- LIVER MASSES

All of the 22 CT patients who had a subsequent diagnostic procedure were submitted either to a repeat CT scan (3 cases), or to one of the two other study procedures, i.e.: nuclear medicine or ultrasound (19 cases) (Table XXVIII).

On the other hand, patients in the NM and US groups were distributed differently according to the type of the first supplementary diagnostic procedure. Among the nine patients in the NM group, two had a repeat scan, two had either a US or a CT scan, four had abdominal or gastro-intestinal series, and one had diagnostic surgery. Among the 36 US patients, 32 had one of the other two competing procedures, two had a non-study (complementary) imaging procedure and two were submitted to biopsy. However, the number of categories presented in Table XXVIII and the small number of subjects present in some of them, does not allow significance testing.

TABLE XXVIII

NATURE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE
AMONG PATIENTS INVESTIGATED FOR LIVER MASS
ACCORDING TO PROTOCOL GROUP

TYPE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE	PROTOCOL GROUP NM	US	CT	TOTAL
Repeat procedure	2	0	3	5
Other study procedure	2	32	19	53
Non-study imaging procedure	4	2	0	6
Diagnostic surgery	1	0	0	1
Biopsy	0	2	0	2
TOTAL	9	36	22	67

Despite the difficulty raised by this somewhat itemized analysis of the nature of first supplementary procedures according to protocol group, it was possible to statistically analyze the type of first supplementary procedures. It was indeed considered useful to find out, when a physician was not entirely satisfied with the results of a given study scan, whether the first supplementary diagnostic procedure performed would more likely be one of the (competing) diagnostic procedures under study or rather a non-competing complementary diagnostic procedure, be it of a radiological or surgical nature. In fact, it assumed that by choosing a supplementary procedure other than one of the three (competing) technologies under study, the physician considered that none of the imaging procedures studied here could provide him with the additional information required, and consequently that none of the competing procedures was diagnostically more useful than the scan already performed. Within this framework, the data presented in Table XXVIII were grouped into two types of supplementary procedures: repeat and competing procedures, the latter being defined as any of the two other diagnostic technologies on which the present study focuses; and complementary procedures, including all other non-study imaging technologies as well as surgical procedures performed for diagnostic purposes.

Results presented in Table XXIX suggest that should a supplementary diagnostic procedure be necessary, patients investigated for a liver mass by nuclear medicine will more likely be submitted to a complementary procedure as a first supplementary test than patients investigated by ultrasound or computed tomography. In fact, five of the nine patients in the NM group for whom supplementary testing was required were submitted to a complementary procedure, as opposed to only 4 of

TABLE XXIX

FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE AMONG PATIENTS INVESTIGATED FOR A LIVER MASS, ACCORDING TO TYPE OF PROCEDURE AND PROTOCOL GROUP

TYPE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE	PROTOCOL GROUP NM	US	CT	TOTAL
Repeat or competing procedure	4	32	22	58
Complementary procedure	5	4	0	9
TOTAL	9	36	22	67

$X^2 = 17.3$; DF = 2; $p < .001$

36 patients in the US group and none of 22 patients in the CT group ($\chi^2 = 17.3$; $p < .001$).

2- PANCREATIC DISORDERS

As in the preceeding section, the distribution of first supplementary diagnostic procedures among patients investigated for pancreatic disorders was tabulated, according to the nature of the (first) supplementary procedure and according to protocol group. Results appearing in Table XXX indicate that 4 of 13 patients investigated by ultrasound and submitted to further testing had a repeat ultrasound scan as the first supplementary diagnostic procedure, while only one of 11 patients investigated by CT and submitted to further testing had a repeat CT scan.

The frequency with which the other competing study procedure was chosen as the first supplementary procedure varies little, from 2 of 13 patients in the US group, to 4 of 11 patients in the CT group. Similarly, non-study imaging procedures were almost as frequent in the US group (7 of 13 patients) as in the CT group (5 of 11 patients).

Only one patient investigated for pancreatic disorder was subsequently submitted to diagnostic surgery; this 52 year old patient was originally investigated by CT; surgery confirmed the CT diagnosis of pancreatic cancer.

As in the case of liver masses, the data appearing in Table XXX were then grouped according to type of first supplementary diagnostic procedure. As indicated in Table XXXI, the frequency of complementary procedures (as

TABLE XXX

NATURE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE
AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS
ACCORDING TO PROTOCOL GROUP

TYPE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE	PROTOCOL GROUP US	CT	TOTAL
Repeat procedure	4	1	5
Other study procedure	2	4	6
Non-study imaging procedure	7	5	12
Diagnostic surgery	-	1	1
Biopsy	-	-	-
TOTAL	13	11	24

TABLE XXXI

FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS, ACCORDING TO TYPE OF PROCEDURE AND PROTOCOL GROUP

TYPE OF FIRST SUPPLEMENTARY DIAGNOSTIC PROCEDURE	PROTOCOL GROUP US	CT	TOTAL
Repeat or competing procedure	6	5	11
Complementary procedure	7	6	13
TOTAL	13	11	24

$X^2 = .001$; DF = 1; $p > .30$

first supplementary diagnostic test) is not significantly different in the ultrasound group (7 of 13 patients had a complementary procedure) than it is in the CT group (6 of 11 patients) (χ^2 = .001; p > .30).

I- MEAN DURATION OF POST-SCAN RADIOLOGICAL AND SURGICAL INVESTIGATION

A final measure of management efficacy relates to the period of time (in days) during which the patient was submitted to diagnostic imaging and surgical procedures for investigation of a given diagnosis after the initial study scan was performed and reported. Given the caveats already outlined (chapter III) concerning the validity and utility of this measure, only the highlights of the analysis are reported here.

1- LIVER MASSES

The distribution of cases according to duration (in days) of post-scan radiological and surgical investigation according to protocol group was tabulated (Table XXXII). Overall, the post-scan duration of investigation varies from zero to 21 days. The mean duration observed for patients investigated by nuclear medicine was 0.79 day (std. deviation = 2.44; variance = 5.96); the mean among ultrasound patients was much higher at 2.09 days (std. deviation = 2.80; variance = 7.86); finally the mean duration among CT patients was 1.57 day (std. deviation = 3.93; variance = 15.47). However, analysis of variance does not show these differences in means to be statistically significant ($F_{2,141}$ = 2.29; p > .05), although they are numerically important.

2- PANCREATIC DISORDERS

Similarly, patients investigated for pancreatic disorders by either

TABLE XXXII

POST-SCAN DURATION OF DIAGNOSTIC IMAGING AND SURGICAL INVESTIGATION FOR LIVER MASSES ACCORDING TO PROTOCOL GROUP

POST-SCAN DURATION OF INVESTIGATION (DAYS)	PROTOCOL GROUP					
	NUCLEAR MEDICINE		ULTRASOUND		CT	
	Frequency	Total	Frequency	Total	Frequency	Total
0	38	0	22	0	26	0
1	5	5	12	1	9	0
2	1	2	3	6	2	4
3	0	0	6	18	0	0
4	1	4	1	4	0	0
5	1	5	1	5	0	0
6	0	0	3	18	2	12
7	0	0	1	7	0	0
8	1	8	4	32	1	8
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	1	11	0	0
12	0	0	0	0	1	12
13	0	0	0	0	0	0
14	1	14	0	0	0	0
21					1	21
TOTAL	48	38	54	113	42	
MEAN	.79		2.09		1.57	

$F = 2.29$; $V_1 = 2$ DF; $V_2 = 141$ DF; $p > .05$

ultrasound or computed tomography were compared for mean duration of post-scan radiological and surgical investigation. The durations varied from zero to 14 days (Table XXXIII). Patients in the ultrasound group had a mean duration of post-scan investigation of 2.37 days compared to a mean of only 1.30 days for patients in the CT group. However, the null hypothesis of no difference between those two means cannot be rejected ($d = 1.266$; $p > .10$).

TABLE XXXIII

POST-SCAN DURATION OF DIAGNOSTIC IMAGING AND SURGICAL INVESTIGATION AMONG PATIENTS INVESTIGATED FOR PANCREATIC DISORDERS ACCORDING TO PROTOCOL GROUP

POST-SCAN DURATION OF INVESTIGATION (DAYS)	PROTOCOL GROUP			
	ULTRASOUND		CT	
	Frequency	Total	Frequency	Total
0	16	0	13	0
1	1	1	4	4
2	1	2	1	2
3	1	3	1	3
4	1	4	2	8
5	3	15	0	0
6	1	6	1	6
7	0	0	1	7
8	1	8	0	0
9	0	0	0	0
10	0	0	0	0
11	1	11	0	0
14	1	14	0	0
TOTAL	27	64	23	30
MEAN	2.37 days		1.30 days	
VARIANCE	14.24		4.31	

$d = 1.266$; $V_1 = 26$ DF; $V_2 = 22$ DF; $p > .10$

Chapter V

DISCUSSION

A- PRINCIPAL FINDINGS - RELATION TO PREVIOUS STUDIES

This section outlines the principal study findings for each chosen indicator of diagnostic and management efficacy, in each of the two major diagnostic groupings under study, and relates these findings to those reported in previous studies.

1- LIVER MASSES

a) DIAGNOSTIC EFFICACY

Results reported in Chapter IV indicate that when used in the investigation of suspected liver masses, CT contributed to a change in diagnostic likelihood in 62% of cases, not significantly different from the figures observed among similar patients investigated by either ultrasound or nuclear medicine even when controlling for technically poor scans and initial (pre-scan) diagnostic likelihood estimates.

Those results compare favourably with previous studies, including that reported by Wittenberg (1980) using an 18-second scanner. If one assumes, in reading Wittenberg's report, that CT contributed to a change in diagnostic likelihood in all cases except those where CT was considered to have had "little or no effect on diagnostic understanding", then Wittenberg's data suggest that CT resulted in change in diagnostic likelihood in 57% of all cases in all protocol groups, although specific data are not available for the liver protocol. Using a different method, Baker and Way (1978) observed that among 202 patients in eight diagnostic categories or indicators, CT provided no new information in 43% of all cases; one could then assume that no change in diagnostic likelihood was observed in at least 43% of patients investigated for various reasons

(only 10% of patients were actually investigated for a liver mass) by the 22-second CT scanner used by Baker and Way.

A comparison with Robbins (1978; 1980) is difficult since he did not separate those cases where CT had no effect on diagnosis from those where it had a negative or misleading effect. Robbins (1978) does report that among the various abdominal applications of the 2.5 minute scanner in his 1978 series, CT provided information not otherwise available and had a positive effect on diagnosis, prognosis or therapy in 23% of cases, and had negative (misleading) or no effect on diagnosis, prognosis or therapy in the remaining 77% of cases. When Robbins (1980) pursued the study with a faster 2-second CT scanner, the corrected results showed little improvement in the proportion of cases for which CT was considered to have had a positive effect on diagnosis, prognosis or therapy. Robbins' data do not allow separation of diagnostic efficacy from management efficacy ratings.

When CT, US and NM were compared for magnitude of change in diagnostic likelihood, measured by a weighted diagnostic efficacy score, the mean score observed among CT subjects was slightly lower than that observed among NM subjects and slightly larger than that among US subjects, although the difference was not found to be statistically significant. No similar measure was used in other studies.

The pre-scan versus post-scan comparison of the proportion of cases considered to be diagnostically definite (very likely or very unlikely) was used as a basis for comparing the contribution of CT, NM and US to improvement of the clinician's diagnostic understanding. Results indi-

cate that all three technologies contributed importantly toward improving diagnostic understanding in this group of patients. The proportion of diagnostically definite cases increased from 36% to 74% among the CT subjects, from 42% to 84% among the NM subjects, and from 44% to 70% in the US group. Furthermore, when the nature of change in diagnostic likelihood estimates was analyzed, such changes were found to constitute improvement in diagnostic understanding in 78% of (27) cases in the NM group, 65% of (23) cases in the US group, and in 73% of (26) cases in the CT group. Between group differences in the increase of diagnostically definite cases and in the proportion of cases where a change in diagnostic likelihood resulted in improving diagnostic understanding were not found to be statistically significant.

Although no other published study has yet compared CT, NM and US for this particular measure of diagnostic efficacy, results for CT patients only are somewhat different from those reported by Wittenberg (1980) who, using an 18-second scanner, observed that CT was considered to have contributed to an improvement in diagnostic understanding in only 47% of patients in the liver protocol. Comparison with other studies (Baker and Way 1978; Robbins 1978, 1980; Dixon 1981) is difficult because of major differences in general methods and in the efficacy measures used.

b) MANAGEMENT EFFICACY

The three technologies were compared on two principal measures of management efficacy, supplementary testing and duration of post-scan radiological and surgical investigation of suspected liver masses.

When technically poor scans were excluded, less than 16% of 45 patients investigated by NM were subsequently submitted to additional diagnostic imaging or surgical procedures, compared to 69% of 51 patients investiaged by US and 55% of 42 patients investigated by CT. These differences, particularly that between the NM group and the other two groups were numerically important and statistically significant.

Because of the fact that the vast majority of initially-requested scans were for nuclear medicine it was very difficult to isolate the effect of the scan on supplementary testing (management efficacy) from the effect of randomization to an other-than-requested scan. However, results of analysis in the US group indicate that supplementary testing is not significantly influenced by concordance to initial scan request. If results of this limited analysis can be applied to the NM and CT groups, and nothing would indicate this to be inappropriate, it is plausible to consider that concordance of scan performed to initially-requested scan is not likely to have had a significant confounding effect on the frequency of patients submitted to supplementary testing.

These results compare only partly with those reported by Dixon et al. (1981). Comparing two groups of patients randomly assigned to CT or to available alternative methods as first procedure in the investigation of palpable abdominal masses, Dixon found that 54% of CT patients (versus 55% of CT liver cases in this series) needed further inpatient investigation as opposed to 76% of patients in the non-CT group, (compared to 16% in the NM group and 69% in the US group in this series). The difference between Dixon's two groups did not quite satisfy the usual requirements for statistical significance (the resulting p value

being .10 > p > .05). It is, however, difficult to compare Dixon's non-CT group to the NM or US group in the present study since Dixon's results obviously reflect the management efficacy not only of nuclear medicine or ultrasound, but also gastroscopy, arteriography, lymphangiography, intravenous cholangiography, oral cholecystography and other procedures to which non-CT patients were first submitted.

Other CT efficacy studies dealing with the issue of supplementary testing have done so by comparing clinicians' plans for further diagnostic procedures before and after the CT scan was performed (Wittenberg 1978, 1980; Robbins 1978, 1980). This important difference in methods makes comparison between those studies and this study, very difficult.

Comparison of the type of first supplementary diagnostic procedure among CT patients to that observed among US and among NM patients investigated for a liver mass, showed that the first supplementary procedure performed was either a repeat of the initial scan or one of the other two (competing) procedures under study in 100% of 22 CT cases and in 89% of 36 US cases, compared to only 44% of nine NM patients. These significant differences suggest that physicians have greater confidence in NM findings.

When the three liver protocol groups were compared on the basis of duration of post-scan diagnostic investigation for the same indication, the observed means were 1.57 days for the CT group, 2.09 days for the US group and 0.79 day for the NM group. However, because of large within-group variations in post-scan durations of diagnostic investigation, the reported differences in means were not found to be statistically signi-

ficant.

No other reported study has used the same measure of management efficacy. However, Dixon (1981) reported on time taken to arrive at a diagnosis among patients investigated for abdominal masses by either CT or available alternative methods. Dixon observed a significantly shorter mean time to diagnosis in the CT group (mean of 9.6 days) than in the non-CT group (mean of 18.7 days). Not only are Dixon's comparative results different from those reported in this series, but also the values obtained for each group are substantially larger (in number of days) than here. Dixon defined time taken to diagnosis as follows:

> "the number of days from the first imaging investigation to a final diagnosis established either by biopsy/laparotomy or by the final imaging investigation; the latter was taken to be the diagnostic end-point when it showed no lesion or unequivocally showed a lesion which did not require immediate surgery or biopsy."

It is likely that this difference in definition or construct of the variable as well as differences in the diagnostic indications for the study scans (liver masses in the present study versus palpable abdominal masses in Dixon's study) contribute to the large difference in the means observed in both studies and make serious comparison of results hazardous. Furthermore as pointed out earlier, the variety and nature of the first imaging procedures in Dixon's non-CT group further compromise the usefulness of comparisons between the two studies.

In conclusion, data on the relative efficacy of NM, US and CT reported in the present study suggest that no important difference in diagnostic

efficacy exists between these three technologies when used in the investigation of suspected hepatic masses. However, comparison of management efficacy indicators suggests that NM is less likely to lead to further testing than US or CT, and that when supplementary tests are performed, patients initially investigated by NM are more likely to be submitted to complementary (non-study) procedures than patients initially investigated by US or CT; this may indicate greater confidence in NM findings.

It may be surprising that NM was shown to have greater diagnostic efficacy than US or CT. Two elements may account, at least in part, for this situation.

When the three technologies were compared for the frequency of supplementary testing, it was not possible to fully analyze the possible effect of concordance to initial scan request, although in the US group where sufficient data were available, no significant effect was observed. Finally, it is possible that small differences in diagnostic efficacy which could not achieve statistical significance given the number of subjects, did contribute to larger and significant differences in management efficacy.

2- PANCREATIC DISORDERS

a) DIAGNOSTIC EFFICACY

CT contributed to a change in diagnostic likelihood in 70% of 23 cases while US contributed to a change in diagnostic likelihood in only 30% of 27 cases investigated for pancreatic disorders, namely pancreatitis

and pancreatic cancer. This difference was found to be statistically significant even when technically poor scans were excluded from the analysis.

Fineberg and Wittenberg (1980) reported that among a set of 73 patients in whom a definitive diagnosis of pancreatic cancer, inflammation or normal pancreas was made (based on autopsy, biopsy or direct surgical observation) physicians' diagnostic estimates were improved in a total of 22 cases. However, physicians were mislead in six cases as a result of the radiologist's interpretation of the pancreatic CT scan. This suggests that physicians' diagnostic probability estimates were changed by the CT scan in 36/37 or 49% of cases, which is substantially less than the results reported here.

However, it should be noted that Fineberg and Wittenberg used a continuous probability scale and did not specify the minimal interval of change; furthermore, this sub-set of 73 patients was composed strictly of patients for whom a definitive diagnosis had been established. Results might have been different had the data reported been based on all patients for whom a CT scan of the pancreas was performed. When magnitude of change in diagnostic likelihood was measured by means of a weighted diagnostic efficacy score, no signficant difference between the mean score of the US group (2.286) versus that of the CT group (2.188) was found. Hence, although CT does contribute to a change in diagnostic likelihood more frequently than does US, the magnitude of that change on the likelihood scale, when it occurs, does not differ significantly from that observed as a result of US.

Improvement in clinicians' diagnostic understanding was measured by comparing the proportion of diagnostically definite cases, at either extreme of the diagnostic likelihood scale, in each group before and after the study scan. Results indicate that the proportion of diagnostically definite cases increased significantly as a result of the study scan only among patients investigated by CT (from 35% to 75% definite cases), while no significant change was observed among the US group (from 67% to 78%).

In cases where changes in diagnostic likelihood were observed subsequent to the initial scan, such changes resulted in improved diagnostic understanding more frequently in the CT than in the ultrasound group.

b) MANAGEMENT EFFICACY

When both groups were compared for the proportion of patients undergoing supplementary diagnostic imaging and surgical procedures, no significant difference was found between the US and the CT group, supplementary testing having been performed in almost 50% of patients in each group.

Similarly no significant difference was found between the type of the first supplementary diagnostic procedure in the CT group and that of the US group. Approximately half of the patients undergoing supplementary testing in each group were submitted either to a repeat scan or to the alternate study scan, as the first supplementary procedure.

Finally, mean duration of post-scan diagnostic investigation was found

not to be significantly different in the CT group (1.30 days) than that observed in the US group (2.37 days), although the CT group did appear to have a numerically shorter mean.

In conclusion, data presented here suggest that when applied in the investigation of pancreatic disorders, CT has greater diagnostic efficacy than ultrasound, although this difference is not reflected in subsequent patient management, as measured here.

B- POSSIBLE LIMITATIONS TO VALIDITY OF RESULTS

As in all field research undertakings, several elements may contribute to influence the validity of the observed results, and should be taken into account when interpreting them.

1- COMPARABILITY OF PATIENT GROUPS

Despite the fact that randomization reduces chances of bias in patient selection and assignment to a given diagnostic procedure (no significant difference in characteristics of patient groups was observed), it is theoretically possible that the indications or reasons for the requested diagnostic procedure included uses other than diagnostic. If this were true, and if such cases were distributed unevenly between protocol groups, findings relating to the diagnostic efficacy of the procedures might have been invalidated by the inclusion of non-diagnostic applications. However, this was not the case since procedures requested for reasons other than diagnostic (based on the information appearing on the requisition form) were systematically excluded from the study by the research and hospital staff. This exclusion procedure was verified on the DAF where respondents were asked to indicate the reason for the

the requested procedure.

Analysis of these data indicate that not one of the procedures was requested for a reason other than to rule in or rule out a diagnosis.

To ensure validity and comparability, eligible subjects for whom any one of the following five conditions applied, were excluded from the analysis:

1- the responsible physician refused to allow the patient's participation
2- the patient refused to undergo the procedure to which he (she) had been randomly assigned
3- the procedure was cancelled by the responsible physician because of a sudden change in the patient's condition
4- the patient was discharged prior to the procedure
5- the patient died prior to the procedure

Patients included in anyone of the five situations listed above might very well have presented characteristics (in the degree of severity of their condition) which could be different from those of other patients in the study population and consequently could have biased the study results.

In those cases where the responsible physician refused to have the patient participate, the possibility exists that such physician would have unconditionally preferred <u>one</u> of the diagnostic imaging procedures under study and therefore bias the results concerning the diagnostic usefulness of the scan.

Patients who refused the procedure could not and were not replaced since, in the two such cases, refusal occurred upon initiation of procedure and

no data concerning the patient's true diagnosis was available at the time. Consequently, there was no way to ascertain whether patients refusing the procedure were actually comparable to those submitting to the procedure. Furthermore, patient acceptability, although not specifically measured here is a relevant issue in the assessment of medical technologies and can in principle constitute a real constraint to their efficacy.

In the case of patients who were discharged or died before the scan was performed, it would seem obvious that the clinical condition of those patients would not have been comparable to those of other patients.

Similarly, patients whose procedure was cancelled by the responsible physician because of a sudden change in the patient's condition could not be considered comparable to other study subjects. In fact, it is highly likely that in such cases, the change in the patient's condition resulted in a more definite diagnosis, probably different from that for which the study procedure had been requested initially.

Thus, the patient groups appear to be comparable and the exclusion of ineligible patients reduced the risk of introducing bias or error.

2- POSSIBLE SOURCES OF ERROR AND BIAS AFFECTING MEASURES OF DIAGNOSTIC EFFICACY

Intra-observer variation (Yerushalmy 1955) in interpretation of computed tomography, ultrasound and nuclear medicine scans was practically non-existent since all scans were interpreted and reported by a single diagnostician in each of the three technologies assessed (except for a

two-week replacement of the CT and ultrasound diagnosticians, during which ten CT study cases and fourteen ultrasound cases were processed). Furthermore, all relevant findings were transmitted to the clinician in written form, according to a standard format. However, no effort was made to systematically standardize the content of the written report for all three technologies. It is possible that systematic intra-observer variation in interpretation capability between technologies was present. If such systematic differences in interpretation criteria did exist between observers in the present study, it would likely be reflected in the number of scans for which the diagnostician refused to identify a specific diagnosis and actually suggested further investigation. When this element was analyzed, the proportion of non-diagnostic scans was 3/65 for CT, 2/77 for US, and 0/48 for NM. It is thus not likely that systematic inter-observer differences in diagnostic interpretation were of importance in this study.

Change in diagnostic likelihood may have been influenced by several factors related either to design of the questionnaire or to methodological design.

The response mode built into the DAF may also have influenced the validity of the results, since by definition, a change in diagnostic likelihood could only be observed when the probability of a diagnosis shifted from one probability interval to another probability interval. It has been shown that it is possible for physicians to make diagnostic probability assessments (Loop and Lusted 1978; Lusted 1977; Wittenberg et al. 1978).

However, the American College of Radiology study (Lusted 1977) and information theory suggest that simple discrete probability intervals constitute a more valid response mode than probabilities recorded on a continuous probability scale since it would appear to be more consistent with actual diagnostic decision-making as well as reducing the effect of various heuristics and biases (anchoring and adjustment, availability, conservatism, scaling effects, etc...) often associated with probability assessment (Tversky and Kahneman 1974).

Use of a continuous probability scale would likely have produced a greater proportion of cases in all groups where diagnostic likelihood was modified as a result of the diagnostic procedure, assuming that all changes (of 1% or more) were considered as such. However, the type of probability scale used in this study offers greater construct validity as well as being more consistent with diagnostic decision-making. This is further demonstrated by the pre-test results. Respondents, having the opportunity of selecting their response mode, overwhelmingly favored the discrete probability intervals as opposed to probability estimates on a continuous scale.

There are bound to be inter-observer variations in diagnostic probability estimations. The diagnostic likelihoods recorded on DAF can only reflect the diagnostic impression of one referring physician at that time, for that patient. Bell (1978) stated:

> "The diagnostic process is a subjective one in the mind of the referring physician. We have no choice but to use a subjective yardstick to measure this process... these data are just as real as more objective data..."

Since the randomization process assured that all patients, regardless of their responsible physician, were equally likely to be submitted to either of the three technologies under study, and since in fact all participating physicians had patients submitted to each of the three modalities, inter-observer variation in diagnostic probability assessment was equally distributed and cannot be considered as a significant source of bias.

Although a continuous probability scale was not used, it is still possible that scaling effects are present in the reported changes in diagnostic likelihood, i.e. the amount of useful information necessary to move diagnostic probabilities from one point on the scale to another may not be the same throughout the scale. Hence, a change from level 1 to level 2 on the DAF may not be as important clinically as a change from level 2 to level 3 and consequently may reflect different diagnostic information content derived by the clinician from a diagnostic procedure. To circumvent this possible shortcoming, changes in likelihood intervals were weighted according to an arbitrary weighting scale developed by consensus of experts and based on the numerical magnitude of the change on the likelihood scale as well as the relative clinical importance of readings at either extreme points on the scale. The initial weighting system, was subsequently submitted to another group of clinicians and was considered to be consistent with the perceived importance of changes in diagnostic likelihood in affecting subsequent patient management. Obviously, a different weighting system based on scientific observation of the relationship between diagnostic likelihood levels and patient management decisions would have had greater validity and may have yielded different results.

Since this weighted value scale was primarily designed to take into account the magnitude and the clinical importance of changes which result in moving diagnostic likelihood toward or away from the two extreme points on the likelihood scale (i.e. very unlikely and very likely), a measure of the overall shift toward or away from the scale extremes might serve to validate results obtained from the analysis of mean diagnostic efficacy scores. Such a measure was in fact developed and reported in Chapter IV in order to compare the improvement in diagnostic understanding resulting from the information provided by each technology.

Thus, when the proportion of total cases at either extremity of the diagnostic likelihood scale recorded before the study scans was compared to the proportion of such cases after the scans, results indicated that all of the technologies (applied to liver masses) contributed importantly and significantly to increase the proportion of "diagnostically certain" cases and that no significant difference in that respect existed between technologies. This concurs with results obtained from the between-group comparison of diagnostic efficacy scores and tend to validate the weighted scale referred to earlier in this section.

Another possible source of bias is the fact that clinicians may have obtained other relevant diagnostic information while waiting for the scan report and that such (other) information may have influenced the post-scan diagnostic assessment. As Bell (1978) pointed out:

> "if a test result is not known for hours or for days, it becomes very difficult for the clinician to isolate just the impact of the one test on his diagnostic thinking."

The mean lag time between performance of the scan and issue of a written report handed to the responsible physician did not vary significantly nor importantly between diagnostic modalities: 1.28 days for CT, 1.23 days for NM and 1.11 days for US. Hence, though it is quite possible that while waiting for the study scan report, clinicians may have been exposed to new information the impact of which has not been isolated, there is no reason to suspect that this possible source of error was greater for any one of the three technologies assessed. Consequently, it is fair to assume that although this shortcoming is real, it is unlikely to have influenced the relative diagnostic efficacy of the technologies being assessed.

Bias might have been introduced by the fact that the scan report was, in each case, handed to the responsible physician by the research staff, and in some cases by the principal investigator.

Simultaneously, the physician was handed the original DAF for the patient and was asked (by the research assistant) to "indicate any change he would like to make to his initial diagnostic likelihood assessment, in view of the scan results." Although in principle the possibility exists that research assistants may have influenced the post-scan diagnostic likelihood, it must be pointed out that all research assistants were strictly directed not to discuss any aspect of the case at any time with the responsible physician. Spot checks were made by the principal investigator to ensure that this procedure was strictly adhered to. However, in retrospect, it might have been wiser to use clerical staff (rather than research assistants with some clinical background) to administer the questionnaire and issue the scan reports, thus eliminating

any chance of bias in this respect.

Some may argue that the change in diagnostic likelihood attributed to the amount of useful information provided by the scan results may be more significantly influenced by the physician's own confidence in the technology and by his knowledge of the technology's diagnostic accuracy in such cases. As argued in Chapter II, clinicians' knowledge of the relative technical efficacy (and diagnostic accuracy) of a technology forms one of the bases from which he will decide to what extent his diagnostic impression will be influenced by the clinical findings of the said diagnostic technology. In the present case, the three technologies have been shown to be of comparable accuracy in the selected applications. We have assumed that knowledge of this fact was evenly distributed through all respondents.

Although this assumption was not tested, it should be pointed out that all respondents were senior residents (and thus had comparable clinical experience) and were exposed to the same clinical environment (Royal Victoria Hospital). Furthermore, it was clearly stated in the project outline given to each participating physician as well as in the verbal briefing each received that the technologies being assessed were known to be equally accurate. None of the participants discussed this issue with the research staff at any time. Hence, although we have not tested respondents' prior knowledge of the relative accuracy of the diagnostic procedures under study, every effort was made to ensure that all respondents were at least aware of the fact that clinical research published up to the time of the study was conducted, indicated that procedures were considered to be equally accurate for the conditions under

investigation in the present study.

Consequently, it is fair to assume that on the basis of their knowledge of the relative accuracy of computed tomography, nuclear medicine and ultrasound, no respondent had reason to have any initial bias toward one technology or another. This is further confirmed by the fact that of the 231 eligible subjects, only one case was excluded because the physician would not allow the patient to participate. If any of the participating physicians was significantly biased toward one technology, one would have expected frequent refusals from these physicians, given that all participating physicians had several of their patients in each of the three technology groups.

3- POSSIBLE SOURCES OF ERROR AND BIAS AFFECTING MEASURES OF MANAGEMENT EFFICACY

Only those diagnostic procedures which according to the information on the test requisition were requested for the same provisional diagnosis as the original scan and performed after the said scan, were actually considered as supplementary. Thus any imaging or surgical procedure designed to investigate a diagnosis other than the one recorded on the initial pre-scan diagnostic assessment was not considered as supplementary, and therefore not recorded for the purpose of the present study.

The validity of this measure may be influenced by the correctness and completeness of the information regarding clinical indications for and date of procedure, recorded by the clinician on the appropriate test requisition form at the time the supplementary diagnostic procedure was

requested. In no instance was the required information incomplete or unclear. In all cases, a brief clinical summary as well as the main reason for the requested (supplementary) procedure appeared on the requisition form, as well as the date the procedure was performed.

Although it was not possible to validate whether all requisition slips for supplementary procedures were included in the medical record of each patient who had been submitted to such a procedure, it was possible to ascertain that the requisition slip for the study scan was present in the medical record of each and every study subject. Since those requisitions were processed by the appropriate clinical department and by the Medical Records department just like any other requisition, it is our impression that under-reporting of supplementary imaging and surgical procedures due to incomplete patient charts was unlikely.

Furthermore, the task of abstracting information regarding frequency and nature of supplementary testing, duration of diagnostic imaging and surgical investigation, was performed according to a pre-established procedure (described in Appendix C-3). Because of the fact that chart abstracting was performed by two persons, inter-observer variation in interpretation of the abstracted information was possible. However, in view of the reliability test reported in Chapter III, the validated abstracting procedure itself would minimize the possibility of significant error or bias of this nature.

As mentioned in Chapter II, the period during which a patient was further investigated by imaging means subsequent to an initial diagnostic (imaging) procedure is a rather indirect or proxy measure of manage-

ment efficacy, and one which, because of uncontrollable confounding variables, can at best serve to confirm results of other measures.

Accordingly, given that, of the patients investigated for a liver mass, the nuclear medicine group was submitted to significantly fewer supplementary procedures, followed by the CT group, then the ultrasound patients with the greatest frequency of supplementary tests, it is not surprising that the mean duration of diagnostic imaging and surgical investigation was shorter for patients in the nuclear medicine group (mean 0.79 day), followed by CT patients (1.60 days) and ultrasound patients (2.06 days), though these differences were not found to be statistically significant. Similarly, it is not surprising that among patients investigated for pancreatic disorders, no important nor significant difference was observed between the two groups (US vs CT) as to the proportion of patients undergoing supplementary tests.

However, it must be remembered that the post-scan period of investigation may be influenced by two other important elements. The first of these is the number of supplementary procedures per patient, which is also a reflection of the efficacy of each of those supplementary procedures.

The average number of supplementary procedures per patient, was not significantly different in the CT group (0.91) than in the US group (0.81) among patients investigated for pancreatic disorders; nor was it significantly different between the nuclear medicine (.31), ultrasound (.94) and CT (.60) liver groups, although the difference is numerically important.

Another source of bias is the nature of the supplementary tests which may influence duration of diagnostic investigation either by variations in waiting times for the procedures (scheduling delays), and in recuperating time necessary before the patient can physically sustain further tests.

It was not possible to control or account for scheduling delays for supplementary procedures nor for prescribed recuperation times relative to each supplementary procedure.

Thus, the absence of any significant difference between the US and CT pancreas groups, as well as between the nuclear medicine, ultrasound, and CT liver groups relative to duration of diagnostic investigation may clearly have been influenced by such uncontrolled variables.

C- LIMITS TO THE GENERALIZATION OF STUDY FINDINGS

Though the present study constitutes the first randomized trial designed to compare the clinical efficacy of computed tomography to that of specific competing technologies in abdominal applications and in that sense may constitute an original and significant contribution to the debate over CT scanning, several limiting factors must be considered with respect to generalization of the findings.

The first of these considerations concerns the indications or applications from which our two major protocol groups were formed. The first group was composed of patients to be investigated by either NM, US or CT for a suspected liver mass during a six-month period, from early

December 1978 to early June 1979. The vast majority (93%) of those patients were actually being investigated for suspected primary or secondary liver neoplasm. The second group was composed of patients investigated by either US or CT for pancreatic disorders, i.e. acute or chronic pancreatitis with or without pseudocyst (34%) and pancreatic cancer (66%), over the same six-month period.

These two groups of patients thus constitute not a sample but the entire population of patients for whom CT, US or NM procedures were requested for the indications specified, over the six-month period at Royal Victoria Hospital.

However, it must be noted that although the clinical applications selected for this study constitute a major portion of those for which CT, US or NM could be requested by Royal Victoria Hospital physicians based on Department of Radiology records, they cannot be considered as necessarily representative of all indications for which any one of these three technologies could be (interchangeably) performed, nor of CT applications in general.

It is important that conclusions drawn from the principal findings of the present study relate strictly to the applications covered by the present research. However, at the time this research was undertaken, the investigation of liver masses and pancreatic disorders constituted (and still does) a sizeable portion of abdominal CT scans and a significant segment of non-neurological CT applications.

A second limiting element is the fact that the investigation was conducted

at a single hospital where diagnosticians and clinicians had a given knowledge of and experience with the three technologies under assessment. Since experience may be an important factor both in the diagnostician's ability to interpret a scan and in the clinician's use of and confidence in the information provided by a given diagnostic technology, and since the Royal Victoria Hospital was not selected as study site on the basis of the representativeness of its clinicians' experience with NM, US and CT, it is possible that results would have been different had participating clinicians and diagnosticians at the chosen study site had more or less experience with these technologies.

Nevertheless it must be stated that clinicians and diagnosticians had been using the CT body scanner for almost 18 months when the study was initiated, in late 1978. Since more than half the CT scanners in use in the United States in 1978 had been acquired during the single year of 1977 (Office of Technology Assessment, 1981) it can be assumed that at the time the present study was initiated, the experience of Royal Victoria Hospital physicians with CT technology was close to the median length of experience of physicians in North America generally. Furthermore, the bulk of hospital-based CT scanners in use in the United States is concentrated in institutions of 500 beds and over, such as Royal Victoria Hospital and more specially in those affiliated with medical schools (Office of Technology Assessment, ibid.). Consequently, although the site for the present research can in no statistical sense be considered as representative of Canadian or American hospitals where CT is used, it can be said that the Royal Victoria Hospital does have the same general characteristics as hospitals where most computed tomography scanners are presently operating.

A third limiting factor to generalization of present findings relates to the role of the technologies under assessment, in the selected clinical applications. The indication for the requested scan, in all cases of any diagnostic group was the first-time investigation of a suspected diagnosis; excluded were other indications such as to guide therapy of an established diagnosis, to provide a basis to assess the efficacy of planned therapy, or to assess prognosis.

Consequently, the relative clinical efficacy of CT in clinical settings where it is used primarily for radiotherapy planning and monitoring may be quite different from that observed here.

The equipment used to perform the NM, US and CT scans reported here must be considered in the context of generalizibility of present findings. All scans were performed on the latest generation of commercially available models for each of the three technologies at the time the study was initiated, i.e. November 1978.

Because the evolution of CT technology particularly has been quite rapid since the first commercial models were sold in the United States in 1974, and because this evolution has continued since the time our study was conducted in 1978-79, the issue of possible obsolescence of the study model must be addressed in regard to present-day generalization of the findings.

However, review of the literature on CT efficacy since 1978 does not establish clearly that technical or clinical efficacy of more recent faster models is superior to that of models such as the one used in the present study.

Robbins (1978, 1980) compared the efficacy ratings obtained with 2.5 minute CT scanner in 1978 to those obtained with an 18-second scanner in 1980, using a similar methodology. After correcting Robbins' data for some reporting errors (see Chapter IV), no difference between the 1978 and the 1980 series was found with regard to the frequency of cases in which CT contributed significantly to an increase in diagnostic information leading to a change in diagnosis, prognosis or therapy, despite a faster scanner and greater reader experience with CT in the later study.

Since the major difference between various generations of scanner machines is faster scanning, one would expect to observe an improvement in image quality and technical efficacy of newer models over older ones, perhaps leading to improved clinical efficacy. However, when various models of CT equipment (including an E.M.I. 5005 model) were assessed for spatial resolution, field uniformity, spatial linearity, artifact resistance and radiation dose using a specially-designed phantom model, and machines were ranked according to results for each parameter tested, Bellon et al. (1979) found that the range of performance or technical efficacy scores exhibited on machines tested was unrelated to class of scanner, or to minimum scanning time which ranged from 3 to 20 seconds, according to class of equipment.

Finally, a senior marketing executive with one of the larger manufacturers of CT confirmed at a national conference on CT (Santé et bien-être Canada, 1978) that in his opinion, the most important improvements in CT technology expected within the next decade or so will be related to improvements in productivity and performance (rather than efficacy) and even doubted that development of new tomography-type technologies,

such as positron emission tomography, would significantly contribute to improved health care services.

In conclusion, there is little evidence to suggest that results of the present study are not generalizable to more recent CT models.

A fourth and final limitation to the generalization of present findings relates to the magnitude of the difference(s) between technologies which could be identified as significant at the .05 level, given the number of subjects in each protocol group (Colton 1974). Comments will be limited to those efficacy indicators and diagnostic groups where no statistically significant difference was observed.

Given the large differences in the costs and potential risks of the three types of procedures under study, it was reasonable to seek large differences between technologies, on the various efficacy indicators. Hence, the number of required cases in each protocol group was estimated on the basis of an expected difference of at least 30% between groups for the various indicators of diagnostic and management efficacy.

The actual sizes of the protocol groups obtained allowed to identify a 30% difference between NM, US and CT (at the .05 level) in the proportion of cases where a change in diagnostic likelihood was observed among liver masses, whereas results showed only a maximum difference of 14% between groups; similarly for results on the diagnostic efficacy score.

Improvement in diagnostic understanding was not found to be significantly different between NM, US and CT when applied to liver masses. Given the

number of actual subjects in each group, only a 40% difference between groups could have been shown to be significant at the .05 level, whereas the largest difference observed was of the order of 15%. In the pancreatic group, at least a 40% difference in the rate of improvement in diagnostic understanding would have been required to achieve significance at the .05 level; results showed a 32% difference between CT and US, at slightly less than the .10 level.

The number of subjects used in the comparison of the various management efficacy indicators only allowed to identify large differences (in the order of 35%) between technologies. When US and CT were compared for such indicators, differences of only 1% or less in the relative frequency and type of supplementary tests were observed, differences largely insufficient to achieve statistical significance, given the number of cases available for analysis.

Hence, when interpreting and generalizing the results of specific analyses showing no significant difference between the technologies under study for a given indicator of clinical efficacy, it must be remembered that given the number of subjects in the study only large differences (of the order of 30% to 40%) could be shown to achieve statistical significance. However, given the large differences in costs between NM, US and CT as well as the differences regarding the safety of these procedures; given that the present study is the first clinical trial designed to compare the clinical efficacy of CT, NM and US; and given the context and means with which the present study was undertaken, it was reasonable and possible to attempt to identify only large differences between the technologies being assessed.

D- POLICY IMPLICATIONS

1- RELEVANCE OF PRESENT STUDY FINDINGS

The possible interest of the findings for policy-makers will depend on the extent to which they contribute to enhance knowledge about various elements considered when making policies about the diffusion and use of new medical technologies.

Typically, such policy will address various concerns such as those found in the first paper on policy implications of the CT scanner published by the Office of Technology Assessment of the U.S. Congress (1976): i.e. development of new technology, its evaluation, its diffusion, its use and financing. Such policies should be based on present knowledge regarding the following seven issues:

a- Technical efficacy
b- Clinical efficacy: diagnostic, management, therapeutic
c- Safety
d- Valuing of technology-derived benefits by recipients of the technology
e- Cost-benefit; cost-effectiveness
f- Importance of clinical applications with respect to major health problems
g- Evaluation of other new competing technologies

Given the definition of technical efficacy outlined in Chapter II and given that the present study was initiated on the basis that the technologies under assessment were of comparable technical efficacy for the clinical applications considered, present study findings obviously do not contribute to enhance our knowledge of the relative technical efficacy of CT and its alternatives.

Clinical efficacy was of major concern in initiating this study which covered various aspects of diagnostic and management efficacy; outcome efficacy was not considered for reasons outlined earlier. In terms of diagnostic efficacy, the findings indicate that when CT is applied for diagnostic purposes to patients investigated for a suspected liver mass, its overall contribution to improved diagnostic understanding is at best not significantly different from that provided by NM or US.

When patient management efficacy measures were considered among patients investigated for a liver mass, the findings indicate that the frequency of supplementary testing subsequent to CT is significantly greater than that observed subsequent to NM; such supplementary tests were more likely to be of a complementary nature in the NM group than tests ordered following a US or CT scan. Duration of post-scan diagnostic investigation did not differ significantly between NM, US and CT patients in this protocol group, although the mean duration observed among NM patients was numerically much smaller than that observed among patients investigated by CT or US.

Thus, the diagnostic efficacy of CT in the investigation of suspected liver masses does not appear to be significantly different from that of ultrasound or nuclear medicine. Furthermore, nuclear medicine appears to result significantly less frequently in additional imaging tests than does CT or ultrasound and when it does, such tests are more likely to be complementary in nature. Thus nuclear medicine, according to the measures used here, does have significantly greater management efficacy than CT or ultrasound.

When similar measures of diagnostic and management efficacy were used to compare CT to US in the investigation of suspected pancreatic disorders, computed tomography was found to have greater diagnostic efficacy than ultrasound on the basis of most of the indicators. It contributed to a change in physicians' diagnostic likelihood estimates more often than ultrasound; it contributed to a significantly greater increase in the proportion of diagnostically definite cases, than did ultrasound; when changes in diagnostic likelihood were observed, CT resulted more frequently to improving diagnostic certainty than did ultrasound.

However, when both technologies were compared for management efficacy, no significant difference was observed beween CT patients and US patients for any of the management efficacy indicators. Thus, although CT seems to improve diagnostic understanding more often and more importantly than US, this advantage is not reflected in the way patients are managed subsequently.

Another important policy issue has not been addressed in the present study: the relative safety of computed tomography. Safety represents a value judgement of the acceptability (for concerned individuals and groups) of the risks associated with the use of a technology. When the safety of a technology is considered, both the nature and probability of the risk(s) should be specified, along with the medical problem to which it is applied, the population affected and the conditions of use of the technology. The Office of Technology Assessment (1981) has integrated these considerations into a definition of risk:

> "A measure of the probability of an adverse or untoward outcome occurring and the severity of the resultant harm to the health of individuals in a defined population, associated with the use of a medical technology applied for a given medical problem under specified conditions of use."

Although not all of these considerations have been covered as yet in the CT literature, some of the published studies have indicated the importance of this issue in the (policy) debate over computed tomography. It has been reported that the amount of radiation to which a patient is exposed during CT examination varies greatly according to conditions of operation (Shope 1980) and machine model (McCullough 1978). Not only should policy-makers be concerned with the short-term consequences of such exposure but also consider the long-term risks as advocated by Fineberg and Hiatt (1979).

The value attributed by patients to the benefits of a technology must be recognized when considering policy for diffusion and use of a technology such as CT and has not been covered in this study. McNeil (1977) illustrated the importance of valuing of health outcomes derived from diagnostic procedures aimed at seeking occult metastases in patients with bronchogenic carcinoma. Although research in the field of "valuing" is still in its infancy, consideration of this concept as it applies to CT in various clinical situations must be an integral part of policy-making considerations in this area. Of particular interest is the value given by consumers to the prolongation of life versus quality of life.

Another element to be considered in policy-making regarding the diffusion and use of new technology relates the net benefits (gross benefits discounted for risks) derived from a particular technology, to the resources consumed in achieving these benefits, in comparison to other alternative

technologies. The present study does increase our knowledge about the relative benefits of CT in comparison to two alternatives. Whenever possible, the benefits should be weighted on the basis of valuing of the result or outcome and only then, related to the costs associated with these benefits. Although cost-benefit and cost-effectiveness studies are complex and are seldom based on complete measures of all benefits and costs, they can help decision-makers set priorities among alternative expenditures for medical care (Fineberg and Hiatt 1979), in this case, in the field of diagnostic imaging.

Results from such cost-benefit or cost-effectiveness analyses should take into account the relative frequency and lethality of the diseases or disease complexes for which the technology is being used. This issue was particularly pertinent in the case of cranial or head CT scanning and paraphrased by Fineberg (1977) as follows:

> "How much are we willing to pay collectively for sophisticated health services of direct benefit to a few?"

In this context, it would seem that policy-makers should relate the cost-effectiveness of this technology for a particular pathology to the cost-effectiveness of other actions aimed at more or less important health problems. Thereby, the relative contribution of a given technology in achieving the goals of the health system can be appreciated and priorities set.

Finally, because of the very rapid evolution of medical technologies as a whole, and of diagnostic imaging in particular in recent years, policy regarding diffusion and use of a technology should take into account the

potential effect of emerging alternatives. In the case of CT scanning, policy-makers would ideally want to avoid the rapid diffusion of a relatively costly technology whose obsolescence in view of developing alternatives such as positron emission tomography (P.E.T.) and nuclear magnetic resonnance (N.M.R.) scanning, might antedate its useful life. As Shroeder and Showstack (1979) pointed out, technology development far outstrips technology evaluation. Thus, policy-makers may want to avoid putting all their eggs in the same basket, given the rapid development of promising new alternatives to CT technology.

2- APPROACH TO RATIONAL DIFFUSION AND USE OF MEDICAL TECHNOLOGY - CT

Given the volume of new technologies being introduced and the complexity of the issues to be tackled, policies regarding the diffusion and use of medical technology in general, and of CT in particular, have raised controversy. On the one hand, there are those who consider that the new technology is diffusing much more widely and rapidly that can be justified on the basis of medical need and knowledge regarding its efficacy. This viewpoint advocates strong government regulation to ensure rational deployment of new scanners and to minimize costs. On the other hand, there are those who consider that the distribution of CT scanners should be determined by the professional judgement of radiologists and other physicians who use them (Relman 1979). The proper approach to this question depends on one's perception of the required nature, content, and scope of policies regarding deployment and use of new technology and on the role accorded to government regulatory bodies and to practitioners in the appropriate planning and use of health resources.

a) NATURE OF POLICY ON DIFFUSION AND USE

Given the nature and complexity of the issues to be addressed by policy-makers regarding new technology, it is unrealistic to expect that information on all of the aforementioned policy items is available. However, the need to explore these issues through ongoing evaluation research is paramount. Such research is considered essential in order to

> "expand our base of knowledge continually, making possible the most informed judgments in the face of remaining uncertainity". (Fineberg 1979, 1)

Thus policy regarding diffusion and use of CT technology must maximize the use of the knowledge base derived both from the manufacturing industry's research and development and by evaluation research. Such policy must reflect the degree of remaining uncertainty, and allow innovation and diversity while recognizing the limits of our knowledge base and the limited availability of financial resources (Shroeder and Showstack 1979).

Many have advocated that diffusion of new medical technology should not be considered until its technical efficacy and safety have been demonstrated (Fineberg and Hiatt 1979, Frazier and Hiatt 1978, Banta and McNeil 1978). In fact, the usual sequence for introducing new diagnostic procedures in medicine begins with their introduction on a limited basis at academic institutions in the context of clinical trials designed to assess efficacy and safety. If the trials are successful, the technology will then be gradually introduced in community hospitals, as experience indicates that the technology does indeed have some benefit and as its acceptance by potential users, increases (Office of Technology Assessment, 1981). Whether this process is allowed to evolve naturally as has been the case with other medical technologies or whether it is enforced by

planning or regulatory bodies is a matter of contention. As pointed out by Lasagna (1978) the protection afforded by the regulatory approach must be balanced against possible social costs resulting from the consequent inhibition of innovation, or the retardation of its spread. Countries with public universal medical and hospital insurance will likely favor a regulatory or planning approach to diffusion of new technology.

However, the issue of diffusion or no diffusion is not the crucial question here.

> "To the extent that a fundamental problem with CT scanning exists, it lies not in the existence of the technology, but in its appropriate use." (Office of Technology Assessment, August 1978)

> "The real question is not whether to allocate expensive and specialized resources like CT scanners, but how to allocate them for optimal use." (Relman 1979)

b) STRATEGIES FOR RATIONAL USE OF CT SCANNERS

A number of propositions have been made for appropriate use of technologies (including CT). Rationing of the total number of (CT) units installed in a given area to serve a given population is considered by many to be a first step toward rational use of this and other new technology (Wittenberg 1980; Fineberg 1979; Relman 1979). As pointed out by Relman (1979):

> "It should be self-evident...That no matter how useful CT scanners may prove to be, they cannot be installed in every town or hamlet in the land nor in every radiologist's private office...There is, in fact, no practical alternative to the regionalization of most kinds of specialized medical facilities."

Rationing through regionalization can provide a context within which the individual physician can judge the value of the possible diagnostic gain in any one patient against the cost of the procedure. As Fineberg (1979) suggests, physicians knowingly deny resources to some patients everyday, but only when resource limits and alternative choices are clear. Rationing on a regional and/or institutional basis can help clarify those resource limits and help clinicians apply more rigorous decision criteria in such circumstances.

The availability of alternative modalities of diagnostic imaging and knowledge of their relative merits in particular applications are important factors in the rational use of computed tomography. Rational choices in the use of this and other competing imaging technologies in specific applications can best be achieved through an organizational framework allowing an integrated approach to diagnostic imaging. It has been suggested (Santé et Bien-être social Canada, 1978) that a restructuring of existing hospital departments of radiology, nuclear medicine, ultrasound, and computed imaging into a single department of diagnostic imaging (with integrated physical lay-out, professional staff and administrative structure) could favor optimal use of each technology by two means. First, such an organization would facilitate the undertaking of pertinent assessment not only of the comparative accuracy of each technology, but also of its comparative effect on diagnostic understanding, patient management and health outcome. Such studies can contribute not only to reduce uncertainty regarding each technology's clinical efficacy among clinicians using it, but also provide criteria for proper patient triaging.

Secondly, such an organization would facilitate the transfer of such knowledge from research to clinical practice and allow the imaging specialist to play an active consulting role in selecting, for the referring physician, the most appropriate imaging technique in a given case.

> "The clinician is in the best position to formulate the patient's diagnostic problems, but the radiologist is in the optimal position to recommend the preferred imaging approach to solving those problems." (Wittenberg 1980, 1)

This general approach would transfer to the staff of the department of diagnostic imaging the task of developing and upgrading their own criteria for appropriate use of CT and other technologies through rigorous evaluation studies and day-to-day confrontation of clinical problems in an organizational context conducive to such undertakings.

This complex on-going task should not be carried out by regulatory agencies. As stated by Shwartz and Joskow (1979):

> "As a practical matter, can an agency define utilization efficiency for a technology that yields widely different benefits from patient to patient?"

On the other hand, regulatory and government bodies do have a responsibility for developing and applying policy regarding the introduction (on an experimental basis) and rational diffusion of such new technology, as well as for funding of on-going medical technology assessment.

It has also been suggested that revision of physician reimbursement mechanisms and physician education (Moloney and Rogers 1979; Shroeder and Showstack 1979; Robbins 1979; Cooper and Gaus 1979) might contribute

to enhance appropriate utilization of expensive medical technologies. However, expansion of our knowledge base and concurrent adjustment of policies remain of prime importance among approaches to rational use of new medical technology.

When a new technology offers the possibility of some marginal improvement in the health of an individual, but at very high cost, we are confronted with the value trade-off of collective resources versus individual benefit. It is imperative that such value judgements with definite societal implication be formed from as complete a knowledge base as possible.

"Society must understand that a major investment is desirable in the development of medical technology assessment and that the cost of funding such research is an integral part of the cost of providing the service." (Fineberg and Hiatt, 1979)

Judgement as to the societal worth of any new medical technology involves the analysis of several complex issues. "It is fruitless for guardians of the public purse and welfare to demand sure and prompt answers to these issues. It is equally shortsighted for physicians to declare efficacy and need unequivocally established when empirical evidence can only support particular levels of efficacy for particular groups of patients at particular costs." (Fineberg and Hiatt, 1979)

It is clear that assessment of CT or of any medical technology will not provide final answers. Findings will always be subject to interpretation; individual values and judgements and practical operational decisions regarding marketing, purchase, diffusion, reimbursement of new technologies

will continue to be made. However, the aim of medical technology assessment is not to achieve a final assessment, but rather to continually expand the knowledge base making possible the most informed judgment in the face of remaining uncertainty by bringing policy and practice into line with knowledge.

In this context, several areas of research regarding the rational use of computed body tomography still require further investigation.

As an initial step, repetition of the present study on a larger scale, i.e. a multi-center collaborative study, would provide a larger patient population, increase the sensitivity of statistical comparisons between technologies, and allow easier generalization of findings. It would be desirable in planning such a collaborative study to ensure that participating centers have the latest commercially available equipment, and that cost components be introduced into the analysis. This could substantially expand our knowledge of the relative cost-effectiveness of CT in specific applications. Similar studies could be repeated for some of the other frequent diagnostic applications of CT, such as lungs, mediastinum, pelvis, retroperitoneum.

Secondly, the net relative benefits of CT will not be measureable until such time as further research is initiated regarding the relative short-term and long term risks of CT (versus nuclear medicine and ultrasound for example). Although CT is generally regarded as a non-invasive technique, the widespread use of contrasting material in CT as well as the injection of radiopharmaceuticals in nuclear medicine applications raise questions about the non-invasive character of these technologies,

and invite further study of the risks (pain and discomfort) associated with their use; obviously, one cannot underestimate the risks associated with repeated exposure to the varying doses of X-radiation resulting from CT scanning.

A third area warranting further research is that of the value given by patients to the relative benefits derived from CT and competing technologies, particularly in the investigation of various forms of cancer for which favorable treatment outcomes are short-term or for which effective treatment is still non-existant (i.e. cancer of the pancreas).

STATEMENT OF ORIGINALITY:

To our knowledge, the present study constitutes the first prospective comparison of computed body tomography to specific alternative imaging technologies for given measures of clinical efficacy.

Through the use of a prospective randomized trial design, it allows for the first time, comparison of the respective contribution of computed tomography, nuclear cedicine and ultrasonography toward improved diagnostic understanding and of the effect of such contribution on subsequent patient management in specific abdominal applications.

Results should prove useful to physicians in proper patient triaging as well as to planners involved in policy-making regarding diffusion and use of computed tomography.

REFERENCES

Abrams H L, McNeil B J. Computed tomography: cost and efficacy implications. AJR 131: 81-7, 1978.

_____, _____. Medical implications of computed tomography (CAT scanning). NEJM Part II, 298: 310-18, 1978.

Adelstein S J, Parker R, Wagner H N. First phase in objective evaluation of new diagnostic tests: application to thyroid uptake of ^{131}I. Investig Radiol 5: 153-63, 1970.

Albert D A. Decision theory in medicne. MMFQ 56: 362-401, 1978.

Ambrose J, Gooding M R, Uttley D. EMI scan in the management of head injuries. Lancet 1: 847-48, 1976.

Arkin H, Colton R R. Tables for Statisticians. Barnes and Noble, New York, 1950.

Armitage P. Statistical Methods in Medical Research. Wiley & Sons, New York. 1971.

Bahr A L, Hodges F J. Efficacy of computed tomography of the head in changing patient care and health costs: a retrospective study. AJR 131: 45-9, 1978.

Baker C, Way L W. Clinical utility of CAT body scans. Amer J Surg 136: 37-44, 1978.

Banta H D, McNeil B J. Evaluation of the CAT scanner and other diagnostic technologies. HCMR: 7-19, Winter 1978.

Barnoon S, Wolfe H. Measuring the Effectiveness of Medical Decisions. Chs C Thomas, Springfield, IL, 1972.

Bartlett J R, Neil-Dwyer G. A clinical study of the EMI scanner: implications for provision of neuroradiological services. BMJ 2: 813-15,

1978.

Béland F. ANOMHI- Un programme en APL pour l'analyse log-linéaire des tableaux de contingence. Québec, ASOPE, Université Laval, 1980.

Bell R B. Efficacy...what's that? Sem Nucl Med 8: 316-23, 1978.

_____, Loop J W. The utility and futility of radiographic skull examinations for trauma. NEJM 284: 236-39, 1971.

Bellon E M, Miraldi F D, Wiesen E J. Performance evaluation of computed tomography scanners using a phantom model. AJR 132: 345-52, 1979.

Cloe L E. Health planning for CT: perspectives and problems. AJR 127: 187-90, 1976.

Cochrane A L. Effectiveness and Efficiency-Random Reflections on Health Services. Nuffield Provincial Hospitals Trust, Abingdon (England), 1972.

Colton T. Statistics in Medicine. Little, Brown, Boston, 1974.

Cooper B S, Gaus C R. Controlling health technology in Medical Technology - the Culprit behind Health Care Costs? Altman S and Blendon R, eds., Washington DC. U S Public Health Service, DHEW publ no (PHS) 79-3216, 1979.

Cotton P B, Denyer M E, Kreel L, Husband J, Meire H B, Lees W. Comparative clinical impact of endoscopic pancreatography, grey-scale ultrasonography, and computed tomography (EMI scanner) in pancreatic disease: preliminary report. Gut 19: 679-84, 1978.

Craddock J M, Flood C R. The distribution of the χ^2 statistic in small contingency tables. Applied Stat 19: 173-81, 1970.

Creditor M C, Garrett R. The information base for diffusion of technology: CT scanning. NEJM 297: 49-52, 1977.

Crighton D. The accuracy and value of cephalo-pelvimetry. J Obs Gyn Br Comm 69: 366-78, 1962.

Dixon A K, Kingham J G C, McLean A M, White F E. Computed Tomography in patients with an abdominal mass: effective and efficient? A controlled trial. Lancet 1: 1199-201, 1981.

Dronfield M W, Ferguson R, McIllmurray M B, Atkinson M, Langman M J S. A prospective randomized study of endoscopy and radiology in acute upper GI tract bleeding. Lancet 1: 1167-69, 1977.

Edwards W. Subjective probabilities inferred from decisions. Psychol Rev 69: 109-35, 1962.

Egdahl R H. CAT scanners and Pheos: diagnostic and policy issues. NEJM 229: 475-6, 1978.

Evens R G, Jost G R. The clinical efficacy and cost analysis of cranial computed tomography and the radionuclide brain scan. Sem. Nucl Med 7: 129-136, 1977.

_____, _____. Utilization of computed tomography units (in installations with greater than one-and-a-half years' experience). Radiol 131: 695-98, 1979.

_____. Economic implications of a new technology: a CT model. AJR 136: 657-65, 1981.

Fagan A. The evolution of CAT scanning in Canada. Dim Health Serv. 54: 49-52, 1977.

Fineberg H V. Computerized tomography: dilemma of health care technology. Pediatrics 59: 147-49, 1977.

_____. Evaluation of computed tomography: achievement and challenge. AJR 131: 1-4, 1978.

_____. Medical technology policies and computed tomography. Ann Int Med 90: 114-15, 1979.

_____, Bauman R, Sosman M. Computerized cranial tomography - effect on diagnostic and therapeutic plans. JAMA 238: 224-27, 1977.

_____, Hiatt H H. Evaluation of medical practices: the case for technology assessment. NEJM 301: 1086-91, 1979.

_____, Wittenberg J. Evaluation of computed tomography in pancreatic cancer. Bull du Cancer (Paris) 67: 390- 94, 1980.

Frazier H S, Hiatt H H. Evaluation of medical practices. Science 200: 875-78, 1978.

French B N, Dublin A B. The value of computerized tomography in the management of 1000 consecutive head injuries. Surg Neurol 7: 171-83, 1977.

George R O, Wagner H N Jr. Ten years of brain tumour scanning at Johns Hopkins, 1962-1972 in Non-Invasive Brain Imaging: Computed Tomography and Radionuclide. Leblanc & Sorenson eds. New York, Society of Nuclear Medicine, 1975.

Haberman S. Analysis of Qualitative Data, vol. 1. Academic Press, New York, 1978.

Hogarth R M. Cognitive processes and the assessment of subjective probability distributions. JASA 70: 271-89, 1975.

Husband J E, Golding S J. Computed tomography of the body: when should it be used? BMJ 284: 4-8 , 1982.

Jacob R. Tomographie axiale par ordinateur-état actuel, document interne, Ministère des Affaires Sociales, Gouvernement du Québec, Mars 1982, unpublished report.

Johsson E, Markė L A. CAT scanners: the Swedish experience. HCMR 2: 37-

55, 1977.

Knaus W A, Schroeder S A. Impact of new technology: the CT scanner. Med Care 15: 533-42, 1977.

_____, Davis D O. Utilization and cost-effectiveness of cranial computed tomography at a university hospital. J Comp Assoc Tomogr 2: 209-14, 1978.

Larntz K. Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. JASA 73: 253-63, 1978.

Larson E B, Omenn G. The impact of computed tomography on the care of patients with suspected brain tumor. Med Care 15: 543-51, 1977.

Lasagna L. The development and regulation of new medications. Science 200: 871-73, 1978.

Ledley R S, Lusted L B. Reasoning foundations of medical diagnosis. Science 130: 9-21, 1959.

Levitt R G, Jost G, Trachtman J, Sagell S S, Stanley R J. A computer-assisted method to determine the diagnostic efficacy of computed tomography of the body. Radiol 123: 97-101, 1977.

Lewontin R C. The robustness of homogeneity tests in 2 x N tables. Biometrics 21: 19-33, 1965.

Loop J W, Lusted L B. American College of Radiology Diagnostic Efficacy Studies. AJR 131: 173-79, 1978.

Lusted L B. Logical analysis in roentgen diagnosis. Radiol 74: 178-92, 1960.

_____. A Study of the Efficacy of Diagnostic Radiologic Procedures. Final report on diagnostic efficacy. Chicago. American College of Radiology, 1977.

McCullough E C. Computed tomography scanner generations in Computed Body Tomography for the Diagnostician, Harvard Medical School, Departments of Radiology and Continuing Education, Boston, 1978.

McNeil B J, Varady P D, Burrows B A, Adelstein S J. Measures of clinical efficacy: cost-effectiveness calculations in the diagnosis and treatment of hypertensive renovascular disease. NEJM 293: 216-21, 1975.

_____, Adelstein S J. Determining the value of diagnostic and screening tests. JNM 17: 439-48, 1976.

_____, Collins J J, Adelstein S J. Rationale for seeking occult metastases in patients with bronchogenic carcinoma. Surg Gyn Obs 144: 389-93, 1977.

_____. Personal communication, September 1978.

Moloney T W, Rogers D E. Medical technology - a different view of the contentious debate over costs. NEJM 301: 1413-19, 1979.

Office of Technology Assessment, United States Congress. The Computed Tomography Scanner and its Implications for Health Policy, Washington, D C, US GPO 1976.

_____. Assessing the Efficacy and Safety of Medical Technologies, Washington, D C, US GPO, 1978.

_____. Policy Implications of the Computed Tomography Scanner: An Update. Background paper, Washington, D C, US GPO, January 1981.

Pauker S G, Kassirer J P. Therapeutic decision-making: a cost-benefit analysis. NEJM 293: 229-234, 1975.

Relman A. CAT scanner - conferring the greatest benefit on mankind. NEJM 301: 1062-63, 1979.

Robbins, A H, Pugatch R D, Gerzof S G, Faling L, Johnson W C, Sewell D H. Observations on the medical efficacy of computed tomography of the chest and abdomen. AJR 131: 15-19, 1978.

_____, et al. Further observations on the medical efficacy of computed tomography of the chest and abdomen. Radiol 137: 719-25, 1980.

Santé et Bien-Etre Canada. Rapport des délibérations du symposium national sur les techniques diagnostiques de visualisation. Ottawa, October 1978.

Schlaifer R. Analysis of Decisions under Uncertainty, New York, McGraw-Hill, 1968.

Schroeder S A, Showstack J A. The dynamics of medical technology use: analysis and policy options in Medical Technology: The Culprit behind Health Care Costs? Altman S and Bleudon R, eds. Washington, D C, US Public Health Service, DHEW Publication no. (PHS) 79-3216, 1979.

Shapiro A R. The evaluation of clinical predictions: a method and initial application. NEJM 296: 1509-14, 1977.

Shope T B. Radiation dose from computed tomograpy: a survey of 20 systems using thermoluninescent and film dosimetry in cylindrical and anthropomorphic phantoms. Manuscript, Bureau of Radiological Health, Food and Drug Administration, DHHS, 1980; referred to in O.T.A., US Congress: Policy Implications of the Computed Tomography Scanner: An Update. Washington D C, US GPO, January 1981.

Schwartz W B, Joskow P L. Should medical costs be curbed at the expense of benefits? Sci Dig: 54-59, August 1979.

Swets J A, Pickett R M, Whitehead S F, Getty D J, Schnur J A, Swets J B, Freeman B A. Assessment of diagnostic technologies. Science 205: 753-59, 1979.

Ter-Pogossian M M. The challenge of computed tomography. AJR 127: 1-2, 1976.

Thorbury J R, Fryback D G, Edward W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiol 114: 561-65, 1975.

Turner D A. An intuitive approach to receiver operating characteristic curve analysis. JNM 19: 213-20, 1978.

Tversky A, Kahneman D. Judgement under uncertainty: heuristics and biases. Sciences 185: 1124-31, 1974.

Welch B L. On the comparison of several mean values: an alternative approach. Biometrika 38: 330-36, 1951.

Wittenberg J, Ferrucci J T Jr. Computed body tomography. Gastroent 74: 287-93, 1978.

_____, _____, Fineberg H V. Evaluating computed tomography. NEJM 298: 1205, 1978.

_____, Fineberg H V, Black E, Kirkpatrick R H, Schaffer D L, Ikeda M K, Ferrucci J T Jr. Clinical efficacy of computed body tomography. AJR 131: 5-14, 1978.

_____. Evaluating efficacy. AJR 134: 1277-79, 1980.

_____, Fineberg H V, Ferrucci J T Jr, Simeone J F, Mueller P R, van Sonnenberg E, Kirkpatrick R H. Clinical efficacy of computed body tomography II. AJR 134: 1111-20, 1980.

World Health Organization. Statistical Indicators for the Planning and Evaluation of Public Health Programs, Geneva, Switzerland. WHO Technical Report Series No. 472, 1971.

Wortzman G, Holgate R C, Morgan P P. Cranial computed tomography: an evaluation of cost-effectiveness. Radiol 117: 75-77, 1975.

Wortzman G, Holgate R C. Reappraisal of the cost-effectiveness of computed tomography in a government-sponsored health care system. Radiol 130: 257-61, 1979.

Yerushalmy J. Reliability of chest radiography in diagnosis of pulmonary lesions. Amer J Surg 89: 231-40, 1955.

APPENDIX A

COMPUTED BODY TOMOGRAPHY STUDY

EXPERT PANEL OF DIAGNOSTICIANS

Composition: Dr. Lawrence Stein
Chief, Department of Diagnostic Radiology

Dr. Robert Lisbona
Chief, Division of Ultrasonography

Dr. Robert Patton
Chief, Division of Nuclear Medicine

Mandate:

1. To outline a protocol for randomization of patients with suspected mass (liver, pancreas) to CT or to its best single alternative, by selecting for each (group of) clinical problem(s), the diagnostic procedure considered to be the best or the first to undertake, in the absence of CT.

2. To cooperate with a group of expert clinicians in the elaboration of a list of possible presenting problems or diagnostic impressions on which referring physicians will be assessed before and after the imaging procedure(s) of interest.

3. To estimate the approximate weekly volume of eligible subjects based on their current experience regarding the number of individual in patients (not number of examinations) submitted to radionuclide scan, to ultrasonography, and/or to computed body tomography for the investigation of liver and of pancreatic masses (including carcinoma suspected on the basis of clinical history).

4. To approve a standard format and standard mode of transmission of imaging results to the referring physicians within the context of the study.

COMPUTED BODY TOMOGRAPHY STUDY

EXPERT PANEL OF CLINICANS

Composition: Dr. Urs Steinbrecher
SNR, Medical Resident

Dr. Marvin Wexler
Department of Surgery

Dr. Peter Mlynaryck
Department of Medicine

Mandate: To assist the principal investigator to:

1. Determine a list of possible diagnoses or diagnostic impressions for patients submitted to computed body tomography, radionuclide scanning, or ultrasonography for the investigation of suspected liver or pancreatic masses (including suspected carcinoma); this list is to be used to assess referring physicians' prior and posterior understanding of the patient's condition.

2. Draft a protocol for the randomization of patients with suspected mass (liver, pancreas) to investigation by either CT or its best (or first) single alternative, by selecting for each group of presenting clinical problems the diagnostic procedure considered to be the best or the first to undertake in the absence of CT.

3. Identify possible treatment categories for the conditions under study; this list of treatment categories will be used to assess retrospectively any difference in treatment approach, between patients investigated by CT and those investigated by alternatives to CT, for specific conditions.

McGill University

Department of Epidemiology and Health

RE: COMPUTED BODY TOMOGRAPHY STUDY

Dear Sir:

Following last Friday's meeting concerning the above subject, I wish to express my appreciation for taking time out from your tight schedule in order to help us clarify some aspects of the project.

Since the discussions were informal, I would appreciate that you go over the summary of decisions that were agreed upon and make any correction or revision you consider appropriate. Please feel free to forward your comments in writing or in person as I would be available for discussing the matter with you anytime at your earliest convenience.

Thank you again for your valuable cooperation.

Pierre Boyle

Department of Epidemiology & Health
392-4743

PB:ml

Encl.

ELIGIBLE SUBJECTS FOR STUDY

It was agreed that eligible subjects would be:

- all in-patients for whom a requisition is received during the course of the study for first-time investigation of some suspected hepatic or pancreatic mass by either radionuclide, CT, or ultrasound imaging
- excluded would be those patients who have been submitted to either one of these procedures for similar reasons during the preceding twelve month period

LIST OF POSSIBLE DIAGNOSES (liver and pancreas)

Once requisitions of eligible subjects have been identified, the participating referrer will be asked to check from a given list of possibilities, the diagnosis(es) which best corresponds to his diagnostic impression of the patient's condition and to indicate his level of confidence about each one. He will be asked to review this choice after the imaging procedure.

Though our discussions have not lead to a clear consensus regarding the list of possible diagnoses, I have drafted one which reflects my understanding of what was said as well as the written suggestions submitted by Dr. Urs Steinbrecher. Please note that subsequent to the opinions expressed during the meeting, we have excluded obstructive jaundice and have concentrated specifically on (suspected) masses.

LIST OF DIAGNOSTIC IMPRESSIONS

LIVER
(suspected masses)

1) CONGENITAL:
   a) accessory lobes _____
   b) cystic disease _____

2) INFECTIOUS:
   a) abscess _____
   b) echinococcus cyst _____

3) NEOPLASTIC:
   a) primary: i = benign _____
      ii = malignant _____
   b) secondary (mestastatic) _____

4) VASCULAR:
   a) hematoma _____
   b) A-V malformation _____

5) OTHER: (specify) _____

6) NORMAL: _____

PANCREAS
(suspected masses)

1) PANCREATITIS:
   a) acute _____
   b) chronic _____
   c) with phlegmon or pseudocyst _____

2) NEOPLASM:
   a) primary: i = benign _____
      ii = malignant _____
   b) secondary _____

3) ABSCESS: _____

4) CYST: _____

5) NORMAL: _____

6) OTHER: (specify) _____

RANDOMIZATION PROTOCOL

After lengthy discussion, it was agreed that randomization of patients would occur after an eligible subject's requisition for any one of the procedures of interest was identified and the referring physician's consent has been obtained.

RANDOMIZATION OF SUSPECTED LIVER MASSES

Thus, all in-patient requisitions for first-time investigation of an hepatic mass by either nuclear medicine (liver-spleen scan), computed tomography, or ultransonography will be randomly assigned (assuming physician's consent) to one of the three procedures (see attached diagram).

RANDOMIZATION OF SUSPECTED PANCREATIC MASSES

All in-patient requisitions for first-time investigation of a pancreatic mass(es) by either CT or ultrasound will be randomly assigned to one of these two procedures (see attached diagram).

It was decided by the group to exclude ERCP from the study, because of ethical considerations.

LIST OF POSSIBLE THERAPEUTIC DECISIONS

Since as a second part to the study we will compare therapy for patients with similar conditions diagnosed through the different procedures under study, it is necessary for us to decide ahead of time the possible therapies which might be recorded from chart abstracting i.e. what kind

of differences in therapeutic approach do we want to pick up? How fine or sensitive should the possible treatment categories be?

I have drafted a tentative list on which you are invited to comment:

1) Surgery
2) Chemotherapy (should nature and dose be considered?)
3) Radiotherapy (same question)
4) Medication
5) Palliative care
6) Order further imaging procedure (specified)

Royal Victoria Hospital

687 PINE AVENUE WEST
MONTREAL, QUEBEC
H3A 1A1

DEPARTMENT OF DIAGNOSTIC RADIOLOGY

HONORARY CONSULTING STAFF
CARLETON B. PEIRCE, M.D.
ROMEO ETHIER, M.D.

ATTENDING STAFF
LAWRENCE A. STEIN, M.D., DIRECTOR

BURT B. HALE, M.D.
ROLLA E. WILSON, M.D.
JEAN H. GAGNON, M.D.
JOSEPH TOTH, M.D.

PATRICK J. FITZGERALD, M.D.
G. BERNARD SKINNER, M.D.
JEAN-MARC DUMAS, M.D.
ROBERT E. HANSON, M.D.
D. ROBERT PATTON, M.D.

SUBDEPARTMENT OF NUCLEAR MEDICINE
ROBERT LISBONA, M.D., PHYSICIAN-IN-CHIEF
B.G. FALLONE, M.SC., PHYSICIST

November 16, 1978

To all members of the Medical and Surgical Staff:

Please take note that for a period of fifteen weeks starting November 20, 1978, the Department of Epidemiology and Health at McGill, in cooperation with the Department of Diagnostic Radiology of the Royal Victoria Hospital, will be involved in a prospective randomized study to assess the relative clinical impact of three diagnostic imaging techniques (nuclear medicine, ultrasound and computed body tomography) on the diagnosis and treatment of jaundice N.Y.D., suspected hepatic masses, and suspected pancreatic masses.

As a result, inpatients for whom a requisition is made for first-time investigation of the above-mentioned conditions by either nuclear medicine, ultrasound or C.T. will be randomly allocated to either of the said procedures, after written consent has been obtained from the senior resident on the ward. Senior residents will be assessed on their pre-scan and post-scan diagnostic confidence, by way of a short form to be attached to the original requisition.

It should be noted that the randomization procedure does not involve any additional risk to the patient and that clinicians will be free to request additional diagnostic procedures following review of the study scan results, if such are considered necessary.

In view of the potential benefits of this study in terms of future development of investigative protocols as well as influencing governmental policy in the field of diagnostic imaging, you are invited to lend your full support to its realization.

L. Stein

Lawrence A. Stein, M.D.

Lloyd D. MacLean, M.D.

M. McGregor

Maurice McGregor, M.D.
Physician-in-Chief

APPENDIX B

DIAGNOSTIC IMAGING STUDY PROCEDURE

FOR SUPPORT PERSONNEL IN

DEPARTMENT OF RADIOLOGY

(Monday to Friday)

1. Please set aside all in-patient requisitions received by your office for examination of the LIVER or PANCREAS.

2. Three or four times daily, Dr. Psihramis or Mr. Pierre Boyle will visit your office. On each occasion the said requisitions will be screened for patient eligibility, the forms* detached; the referring physician will be contacted by telephone, and the patient randomized to investigation by either nuclear medicine, ultrasonography or computed body tomography.

3. You will then be informed immediately of any changes to be made to the original requisition(s). Modified requisitions should then be dispatched to the appropriate office:

   Room 204 for nuclear medicine

   Room 511 for ultrasound

   Room 526 for computed body tomography

4. Patients will then be scheduled as per routine for the imaging procedure to whichthey were allocated.

5. Reports of all nuclear medicine, ultrasound, and CT examinations of the liver and pancreas should be held until 16h (4 p.m.) daily at which time Mr. Boyle or Dr. Psihramis will screen them and identify reports pertaining to study subjects.. The name and telephone number of the referring physician will then be recorded by the researcher who will attempt to interview the referring physician as soon as he has received the (verbal or) written report of the examination.

* You will notice that a yellow "DIAGNOSTIC ASSESSMENT FORM" will accompany many of these requisitions. In principle all requisitions pertaining to eligible study subjects should be accompanied by this form.

PIERRE BOYLE

Department of Epidemiology
392-4743

DATA COLLECTION PROCEDURE

AND DAILY SCHEDULE

9h30 - 11h

1. Visit secretarial offices of

   First: Dr. Lisbona - 2nd Med.: Angie and Mirca
   Second: Dr. Patton - 4th Surg.: Kathy Dwyer
   Third: Dr. Stein - 4th Surg.: (Miss) Mabel Beighton

   for requisitions of eligible patients (LIVER and PANCREAS)

2. Verify that provisional diagnosis pertains to one of those appearing on D.A.F.

3. Ascertain that:

   - patient has not had a previous NM, US, or CT scan for same problem in last 12 months (either from requisition or from NM, US, or CT records or from the resident).
   - patient is 18 years of age or older
   - for females: should be either over 40 years
     or have had sterilization procedure
     or hysterectomy
     or that original request was for CT scan

4. Attribute a random number of each case in sequential order making sure that patient is entered into the appropriate randomization protocol:

   - liver masses and jaundice, NYD < 6 Bil.
   - pancreatic masses
   - jaundice NYD > 5 Bil.
   - subhepatic abscess

5. Enter following information in upper right-hand corner of D.A.F.

   - patient name and surname
   - location
   - patient number
   - date of birth
   - original test requested and study nos.
   - name of staff man and resident
   - date of scan

6. Make a new requisition when necessary. Note that ultrasound requisitions are also used for CT: in such cases write CT SCAN beside name of organ to be scanned.

7. Draw a red dot on the first and last copies of the requisition.

12h - MEDICAL FLOORS: (note: Medical rounds 10h-12h noon)

Meet with senior resident to inform him that we have entered patient "X" into our study protocol and that we would like him to fill in the D.A.F.

Note that likelihood levels should not total > 6. Inform resident of test to which patient was allocated and when it will be done, making sure that he signs the consent form before.

SURGICAL FLOORS:

Same procedure but since surgical residents are in the O.R. every day until about 4h p.m. (except Thursday when they are in oncology clinic) try to reach them by phone on the ward or through locating (611) from about 3h30 p.m. onwards. Do not leave until they have all been seen.

P.S. ULTRASOUND PATIENTS N.P.O. 12 hours

2h30 - 3h30
- check the offices for reports of scans done the previous day
- enter results on D.A.F. as follows: +, -, or U (for "uninterpretable").
- take report with you to the ward
- show the report to the resident and ask him if he would like to make any revision to his original diagnostic assessment (making sure you show him the original D.A.F.) on the basis of the scan report
- make sure no verbal reports given; if so, interview the resident on same day
- make sure that at least a preliminary written report is available on CT cases not later than 24 hours after the scan. These CT reports can usually be found either on Dr. Stein's desk or on Miss Beighton's desk.

4h15 Call secretarial offices of any new cases

NOTE: 1. If a case is cancelled, find out why
A- If reason is that another test was done in the meantime (i.e. biopsy or ERCP) then do NOT replace patient by a new one.

B- If patient died before scan, DO NOT REPLACE

1. Keep D.A.F. of cases cancelled and write reason for it.

Finally, when post-scan assessment has been done, enter data on patient data sheet and code according to coding sheet. Attach D.A.F. to completed patient data sheet and store.

Indicate scan results and any undue delay of > 24 hours in comments section of patient data sheet. Also include in comments, any report of pathology found outside suspected area of disease or any unsuspected diagnosis.

APPENDIX C

FOR SUSPECTED HEPATIC & PANCREATIC MASSES

CHECK (V) REASON FOR ORDERING THIS TEST (CHOSE ONE)

____ TO RULE IN OR CONFIRM A DIAGNOSIS

____ TO RULE OUT A DIAGNOSIS

____ TO GUIDE THERAPY OF AN ESTABLISHED DIAGNOSIS

____ TO HAVE BASIS TO ASSESS EFFICACY OF PLANNED THERAPY

____ TO ASSESS PROGNOSIS

DIAGNOSTIC POSSIBILITIES	LIKELIHOOD OF PRESENCE OF THIS DISEASE
INSTRUCTIONS: PLEASE SELECT THE DIAGNOSIS(ES) FOR WHICH THIS PATIENT IS BEING INVESTIGATED, LIMITING YOUR CHOICE TO NO MORE THAN THREE DIAGNOSTIC POSSIBILITIES, AND INDICATE THE LIKELIHOOD THAT THE CHOSEN DIAGNOSIS(ES) IS(ARE) PRESENT.	RATE LIKELIHOOD from 1 to 4 : # 1- very unlikely (<25% probability) # 2- unlikely (25-49% probability) # 3- likely (50-74% probability) # 4- very likely (75% + probability)

LIVER

MASSES-Infectious-Abscess-Intrahepatic ..... #_____
-Extrahepatic-subhepatic ..... #_____
-subphrenic ..... #_____
-Echinococcal cyst ..... #_____
-Neoplastic-Primary benign ..... #_____
-Primary malignant ..... #_____
-Secondary (metastatic) ..... #_____
-Vascular -Hematoma ..... #_____
-A-V Malformation ..... #_____
-Other (specify) _____ ..... #_____

JAUNDICE N.Y.D. -Obstructive ..... #_____
-Non-obstructive ..... #_____
#

NORMAL ..... #_____

PANCREAS

MASSES-Pancreatitis-Acute ..... #_____
-Chronic ..... #_____
-With pseudocyst ..... #_____
-With abscess ..... #_____
-Neoplastic -Primary benign ..... #_____
-Primary malignant ..... #_____
-Secondary (metastatic) ..... #_____
-Other (specify) _____ ..... #_____

NORMAL ..... #_____

Having been informed of the purpose and nature of the above-titled study, I agree to participate and hereby give consent to the random allocation of this patient to either ultrasound, computed tomography, or radionuclide scanning (the latter excepted in pancreatic evaluation). I understan that such randomization will not involve any additional risk for the patient, and that I am free to request any supplementary investigation considered necessary following review of the study scan results.

staffman's signature

senior resident's signature

CODING MANUAL

COL. NO.	INFORMATION	CODES	SOURCE
1-3	SUBJECT NUMBER	101-699 Liver mass jaund 6 Bil. nos. ending 1,2,3 - N.M. " " 4,5,6 - U.S. " " 7,8,9 - C.T.	Table Rand. Nos
		701-799 Jaundice 5 Bil. Even - U.S. Odd - C.T.	Table Rand. Nos
		801-899 Subhepatic mass Even - U.S. Odd - C.T.	" " "
		901-999 Pancreatic mass Even - U.S. Odd - C.T.	" " "
4	PATIENT'S WARD	1- Med. 2- Surg.	Requisition for U.S., C.T. or N.M.
5-6	PATIENT'S AGE	Read as is (in years)	Patient I.D. Card on D.A.F.
7	PATIENT'S SEX	1- Female 2- Male	" " "
8	PROCEDURE REQUESTED	1- Liver-spleen scan 2- Liver-lung scan 3- HYDA scan 4- U.S. 5- C.T.	Requisition
9	REASON FOR PROCEDURE	1- Rule in or confirm D_X 2- Rule out D_X 3- Guide R_X of established D_X 4- Have basis for assessing planned R_X 5- Assess prognosis	D.A.F.
10	PROCEDURE PERFORMED	1- Liver-spleen scan 2- Liver-lung scan 3- HYDA 4- U.S. 5- C.T.	
11-12	LAG TIME	01: 24 hrs 02: 24-48 hrs 03: 48 hrs	
13-14	DIAGNOSIS 1 (D_1) Highest likelihood		Diagnostic assessment form

CODING MANUAL

COL. NO.	INFORMATION	CODES	SOURCE
15	PRIOR LIKELIHOOD OF D_1	DIAGNOSTIC CODES:	Diagnostic assessment form
16	POSTER. LIKELIHOOD OF D_1	01 Liver. Intrahepatic absess 02 - Extrahepatic absess subhep. 03 - " " subphren.	" "
17-18	DIAGNOSIS 2 (D_2) second highest likelihood	04 - Echinococcal Cyst 05 - Primary benign neoplasm 06 - Primary malignant neoplasm 07 - Secondary neoplasm 08 - Hematoma	" "
19	PRIOR LIKELIHOOD OF D_2	09 - A-V Malformation 10 - Other INFARCT 11 - Other	" "
20	POSTER. LIKELIHOOD OF D_2	12 - Jaundice NYD - Obstruct. 13 - " " Non-obstruct. 14 - Normal	" "
21-22	DIAGNOSIS 3 (D_3) Third Highest Likelihood	15 Pancreas Acute Pancreatitis 16 - Chronic Pancreatitis 17 - Pancreatitis c pseudocyst 18 - " c abscess 19 - Primary benign neoplasm	" "
23	PRIOR LIKELIHOOD OF D_3	20 - Primary malignant neoplasm 21 - Secondary neoplasm 22 - Other DIRECT INVASION	" "
24	POSTER. LIKELIHOOD OF D_3	23 - Other 24 - Normal	" "
		LIKELIHOOD CODES 1- Very Unlikely (25%) 2- Unlikely (25-49%) 3- Likely (50-74%) 4- Very Likely (75%)	

CODING MANUAL

COL. NO.	INFORMATION	CODES
25-27	NATURE OF DIAGNOSTIC DECISIONS (POST SCAN)	
25	Supplementary Radioisotope scan	1- No 2- Yes
26-27	Days post-scan of procedure	(numerical)
28	Supplementary C.T. scan	1- No 2- Yes
29-30	Days post-scan of procedure	(numerical)
31	Supplementary U.S. scan	1- No 2- Yes
32-33	Days post-scan of procedure	(numerical)
37	Other Supplementary Non-invasive (upper G.I. series)	1- No 2- Yes
38-39	Days post-scan of procedure	(numerical)
40	Post-scan Biopsy	1- No 2- Yes
41-42	Days post-scan of procedure	(numerical)
43	Post-scan E.R.C.P.	1- No 2- Yes
44-45	Days post-scan of procedure	(numerical)
46	Post-scan Transhepatic Cholan.	1- No 2- Yes
47-48	Days post-scan of procedure	(numerical)
49	Post-scan Arteriography	1- No 2- Yes
50-51	Days post-scan of procedure	(numerical)
52	Post-scan Diagnostic Surgery	1- No 2- Yes
53-54	Days Post-scan of procedure	(numerical)
55	Other Post-scan Diagnostic Decisions (abdominal series)	1- No 2- Yes
56-57	Days Post-scan of procedure	(numerical)

SEE ABSTRACTING PROCEDURE

CODING MANUAL

COL. NO.	INFORMATION	CODES
58-59	Post-scan length of stay	(numerical two-digit)
60-61	Post-scan length of Radiological Work-up	(Numerical two-digit)
62	Ascertainment of Discharge Diagnosis	1- surgery 2- biopsy 3- autopsy 4- other 5- not ascertained at discharge
63	Accuracy of Imaging Scan	1- true positive 2- false positive 3- true negative 4- false negative 5- uninterpretable scan 6- discharge Dx not ascertained 7- non-committent scan

SEE ABSTRACTING PROCEDURE

CHART ABSTRACTION PROCEDURE

REQUIRED INFORMATION	SPECIFICATION	SOURCE OF REFERENCE
Post-scan diagnostic imaging and surgical procedures as listed below:	All procedures listed below which were performed after date of study scan, for similar indication and episode of illness defined as less than one month after initial study scan with same suspected pathology; when date of such procedure is same as that of study scan, consider as supplementary only those where clinician indicated he would still have ordered it, as per information in top right corner of D.A.F. Il mention of supplementary diagnostic procedure (as listed) is found in discharge summary but no report available see "progress notes" for date of supplementary procedure; il no mention of supplementary procedure in discharge summary and no report, then consider no supplementary procedure performed.	Date of study scan - N.M., U.S., C.B.T. report - D.A.F. Study scan indication: D.A.F. (D_1) and N.M., C.B.T. or U.S. report.
Radioisotope liver scan	as above	nuclear medicine report discharge summary/progress notes
C.B.T. scan	"	C.B.T. report discharge summary/progress notes
Ultrasound scan	"	Ultrasound report discharge summary/progress notes
I.V. cholangiogram	"	I.V.C. report discharge summary/progress notes
Other non-invasive imagins	upper G.I. series only, other specs. as above	U.G.I. report discharge summary/progress notes
Post-scan biopsy	non-operative (needle) biopsy only other specs. as above	pathology-cytology report pathology consult sheet discharge summary/progress notes
Post-scan E.R.C.P.	endoscopic retrograde cholongiopancreatography other specs. as above	E.R.C.P.-Radiology report discharge summary/progress notes
post-scan transhepatic cholangiography	specs. as above	T.H.C.-Radiology report discharge summary/progress notes
post-scan arteriography	" " "	artériography/Radiology repo discharge summary/progress notes

CHART ABSTRACTION PROCEDURE

REQUIRED INFORMATION	SPECIFICATION	SOURCE OF REFERENCE
Post-scan diagnostic surgery	Surgery considered diagnostic when: - described as "exploratory" - " " laparotomy - mention that not all pathological tissue removed at operation	Surgical report discharge summary/progress notes
Other post-scan D_x decision	Abdominal series only Other specs. as above	Radiology report discharge summary/progress notes
Date of discharge (or death)		discharge summary
P.S. length of stay	Difference in days between date of study scan report and date of discharge (or death)	N.M., U.S. or C.B.T. report discharge summary
Primary discharge diagnosis	Diagnosis relevant to (suspected) pathology for which study scan was performed (D.A.F.) when such is different from primary discharge diagnosis.	Discharge summary Surgical report Pathology report Other Radiology report Autopsy report Progress notes
Ascertainment of discharge diagnosis	Definite confirmation of relevant discharge diagnosis; when no definite confirmation of diagnosis from anyone of the three specified sources, diagnosis considered not ascertained at time of discharge.	Surgical report Pathology report (biopsy) Autopsy report
Accuracy of imaging scan	True positive: when study scan report indicates presence of suspected pathology which is later ascertained by surgical report, pathology report, or autopsy. False positive: positive finding on study scan report with negative finding according to one of the three aforementioned sources. True negative: negative scan report with negative clinical finding according to one of the 3 aforementioned sources. False negative: negative scan report with positive finding from one to the 3 sources. Uninterpretable scan: when recorded as such in top right section of D.A.F. or when the expressions "poor uptake" or "technically poor scans" appear on the appropriate N.M., U.S. or C.B.T. report. Non-committent scan: when appropriate N.M., U.S. or C.B.T. report does not state clearly the interpretation given to the scan and suggests further evaluation. Discharge diagnosis not ascertained: when relevant discharge diagnosis was not infirmed or confirmed by surgical, pathological or autopsy report, in cases where study scan was technically satisfactory and where a clear diagnostic interpretation appeared on the scan report (N.M., U.S. or C.B.T.).	