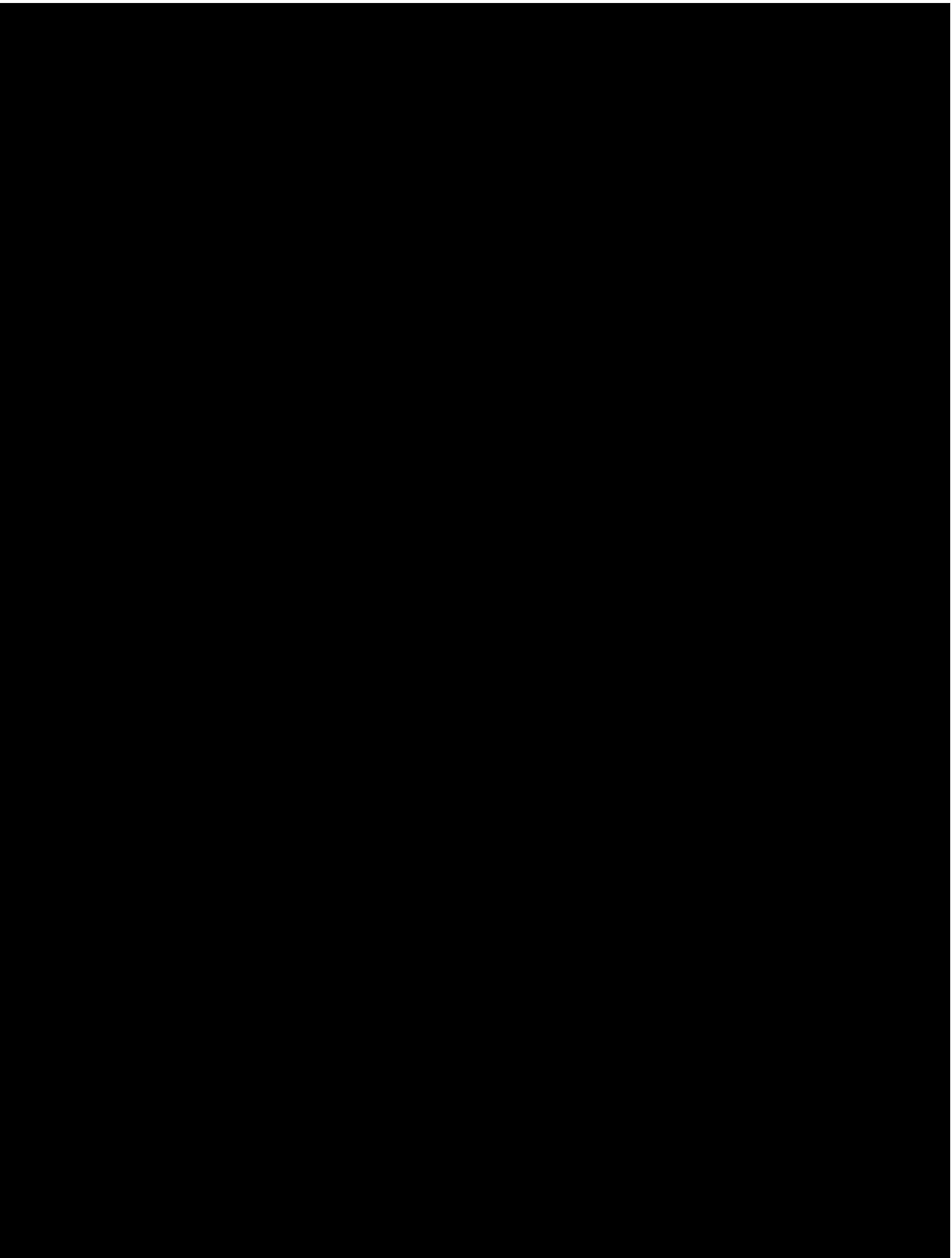


This blank page was inserted to preserve pagination.

MIT/LCS/TR-169

DIAGNOSTIC PLANNING AND CANCER MANAGEMENT

by

Charles Safran

Jane F. Desforges

Philip N. Tsiohlis

September 1976

Massachusetts Institute of Technology
Laboratory for Computer Science,
Clinical Decision Making Group,
Cambridge, Massachusetts 02139.
and
The Blood Reasearch Laboratory,
Tufts New England Medical Center Hospital,
Department of Medicine,
Tufts University School of Medicine,
Boston, Massachusetts 02111.

This research was supported in part by the Health Resources Administration, U.S. Public Health Service under grant number MB 0017 from the Bureau of Health Manpower, and by U.S. Public Health Service Grant number CA 19122 from the National Cancer Institute.

CONTENTS

I. Preface 1
II. Introduction 5
A. State of the art cancer management 5
B. The problem in Hodgkin's disease 7
C. Goals and motivation 8
III. A prototypical System 11
A. Glossary of Terms 11
B. An example session 12
C. Design and structure of components 25
1. The data base 27
i. Data format 29
ii. Searching the data base 31
iii. Prognostograms 34
iv. Problems with the data base and data classifications 36
2. Bayesian Estimation of Tumor Extent 37
i. Bayes' theorem and medical diagnosis 38
ii. Bayes' theorem for Hodgkin's disease 40
iii. Sequential estimation of tumor extent 41
iv. Evaluation of test results 42
3. Decision Analysis methodology 45
i. Decision Trees 46
ii. Utilities 55
iii. Processing of a decision tree 57
iv. Thresholds 58
v. Formation of a Diagnostic plan 59
vi. Sensitivity analysis 59
4. Computer implementation 60
i. Dynamic programming and recursive control structure 61
ii. Building a patient model 62
iii. Control of information flow 63
IV. Results 64
A. Educational benefits 64
B. Verification of Bayesian techniques 66
C. The role of specific tests 67
V. Future directions 70
VI. Tables 73
VII. Appendix I - Staging Classification 78
VIII. References 81

I) Preface

This report describes a computer system for evaluating patients with Hodgkin's disease which has been deveioped by the Clinical Decision Making Group (CDMG) at MIT's Laboratory for Computer Science in conjunction with the Blood Research Laboratory of the New England Medical Center Hospitals and the Department of Hematology, Tufts University School of Medicine (T-NEMCH). This system uses decision theoretic techniques to aid in the formulation of a diagnostic plan for the cancer patient. During the process of plan formation, a patient model is constructed and modified to help suggest and evaluate reasonable diagnostic and therapeutic alternatives. The system also directs sensitivity analyses to determine the effect of error on the decision making process.

Designing this system to operate in a clinical setting to provide expert analysis of complex staging decisions, has required extensive interaction between the physicians at T-NEMCH and members of CDMG so that many people have contributed to the system's development. Professor G. Anthony Gorry, currently at Baylor College of Medicine, formed the CDMG and guided this group in their atempt to developing computer programs which display expert medical decision making skills. His work with Dr. William B. Schwartz on the use of decision theoretic techniques to model certain elements of clinical judgment, laid a foundation upon which the current system was built. During a medical grand rounds, Dr. Jane Desforges, senior hematologist at the New England Medical Center Hospitals and a Professor of Medicine at Tufts University Medical School, presented a discussion of the staging of patients with Hodgkin's disease. After a detailed description of the costs and benefits of diagnostic procedures used to stage patients with Hodgkin's disease, Dr.

Schwartz suggested to Dr. Desforges that decision theoretic techniques might be helpful in evaluating these complex staging decisions. Thus the idea for this project was born. The prototypical development of this computer system was sponsored at the Laboratory for Computer Science by Health Resource Administration, U.S. Public Health Service, Bureau of Health Manpower. Currently, the National Cancer Institute, Division of Cancer Control is supporting this computer system's evolution into a tool for the cancer specialist.

Drs. Jane F. Desforges is principal investigator of the Hodgkin's disease project. Dr. Desforges along with Drs. Philip N. Tsichlis and Avrum Z.Bluming have been responsible for the program's medical expertise and knowledge base. Dr. Stephen G. Pauker, cardiologist and computer extraordinare has always asked the tough critical questions. Without his insight and other invaluable contributions our work would be embryonic instead of infant. Profs. William A. Martin and Peter Szolovits at MIT and Drs. Robert S. Schwartz and Jerome P. Kassirer at T-NEMCH have also provided insightful guidance and criticism.

The actual design and implementation of the Hodgkin's disease system were synthesized from ideas and computer programs developed by members of CDMG. Clearly, the **Digatalis Therapy Advisor** developed by Howie Silverman provided a strong example of a successful working medical advisory computer system which modified quantitative techniques with patient specific knowledge. Kirk Denicoff has contributed many long hours to the systems development. Other members of the group past and present and members of the Laboratory for Computer Science: Byron Davies, Peter Miller, Ramesh Patil, Rand Krumland, Mike Genesereth, Brian Smith, Ken Kahn, and Robert Frankston

have written code and contributed ideas too numerous to mention. As our project became involved with various aspects of data collection and analysis many statistical question arose. Professor Arnold Barnett from MIT's Sloan School of Management with the help of Ellen Eisen is currently studing many of these statistical questions. Finally, one of the coinvestigators, Charles Safran, is leaving the secure computer world for the unknown perils that await him as a first year medical student at Tufts Medical School. Byron Davies will be responsible for the computer system's development during the forthcoming year.

Besides the interaction of people at MIT and T-NEMCH, experts in the treatment of Hodgkin's disease have provided access to their patient records so that the system might be calibrated on the best available information. Dr. Henry S. Kaplan of the Stanford University Medical School, whose pioneering research helped develop a cure for the once incurable Hodgkin's disease, generously provided his computer data base of 909 patients with Hodgkin's disease. These patient records represent the best controlled and best studied group of Hodgkin's disease data in the world. Dr. Samuel Hellman, of the Havard University School of Medicine also provided access to the radiotherapy records of the Peter Bent Brigham Hospital. In addition, many other experts have taken the time to answer difficult questionares.

This report presents 1) the working system being used at T-NEMCH, 2) results of analysis in Hodgkin's disease, 3) the design of a decision support system, and 4) some ideas for future development. Since our work touches areas in computer science, management science, and medicine, many readers may want to selectively read several of chapters. The beginning sections of this report describes some of the problems which physicians face

when managing cancer patients. Although this medical description can be easily glossed over, it has been included to provide some motivation for the many design decisions that have been made in our current implementation. After the description of the problems in cancer management, and specificly in the management of a patient with Hodgkin's disease, a working prototype is discussed which is currently in use at T-NEMCH. While the model which is used in the prototype is tailored for each patient, we have also used the model to yield some interesting medical results. These results, which might only interest those with a medical orientation, are included for completeness and to indicate the broad significance of this research.

The purpose of the report is to present a model for the management of cancer patients which helps elucidate how differing diagnostic and strategies arise, and how some differences can be resolved by a systematic approach to the cancer patient. This model provides a framework for cancer management which includes diagnostic evaluation and treatment selection for a particular patient, the design and evaluation of clinical protocols, and the storage and use of accumulating patient data. The model presented in this report is not meant to duplicate the cognitive style of physicians but to improve on this style in two important areas, the evaluation of uncertainty, and the integration of diagnostic "lookahead" into the decision making process. It is hoped that the use of such a model will extend the expertise of the practicing physician and provide a framework for better decision making. portion Introduction

II) Introduction

A) State of the art cancer management

Cancer broadly defines a disease process which will kill if untreated by continuing growth and extension in the patient. In recent years studies from the National Bureau of Health Statistics have reported increased rate of death attributed to cancer. This increased incidence has also been associated with an increase in national awareness and dedication to find cures from a disease state to which none of us is immune. Perhaps our national preoccupation with cancer has forced upon society an unwarranted allocation of resource when one considers other basic health care needs. The "state of the art" cancer management in our bicentennial year can be roughly characterized by, "early detection" and "aggressive treatment." An explosion of diagnostic technology has become available to the clinician for early detection of malignant and premalignant states. As each new technique is added to the physician's diagnostic armamentarium the costs of early detection or screening increase without evidence of a corresponding increase in patient survival and well being. Furthermore, new therapies which extend symptom free survival have hidden risks of their own that will only become apparent in time. Within the last 10 years, some cancers like Hodgkin's disease have improved survival rates, while others like lung cancer have not improved inspite of early detection and more aggressive therapy.

Decision making in uncertain environments seems to be a predominant feature in the management of many kinds of malignancies. There is basic uncertainty about the biologic phenomenon called cancer, uncertainty about the significance of earlier obtained patient information, uncertainty about the diagnostic accuracy of the tests used to evaluate the

tumor extent before treatment selection, and uncertainty about appropriate treatment. The interpretation and integration of these four kinds of uncertainty into a comprehensive diagnostic and therapeutic plan is the essence of expert clinical judgment.

Uncertainty about the causes of cancer and about the growth properties of the specific tumors lie at the root of the treatment problem. The extent of tumor involvement is a very important determinant because depending also on the biologic characteristics of the particular tumor this can direct the physician to employ different treatment modalities: local excision, radiotherapy, or chemotherapy.

Complex clinical judgments made by cancer experts range from the initial diagnostic evaluation and treatment selection to the long term follow up management of the patient. To compound the complexity of the decision process, many experts may be needed to perform different phases of the management process. Since early detection may offer some cancer victims a chance of "cure," the initial evaluation and treatment plan are a key concern. If one visits many of the nation's leading cancer treatment centers, one finds a number of different diagnostic strategies that are employed on similar patients. Some of these different diagnostic approaches lead to differing treatment selection. Since most patients are not seen at major treatment centers, one suspects that the variance between strategies may be great. Two important questions arise because of this discrepancy: 1) Which of the strategies offers the best expected survival for the patient? 2) If many of the strategies offer comparable results in terms of survival, which strategies optimize other factors such as the quality of life or resource expenditure?

B) The problem in Hodgkin's disease

Within the past decade, pioneering human experimentation has lead to dramatically improved survival for patients with Hodgkin's disease, a malignancy of the lymphatic system. The selection of an appropriate therapy, either megavoltage radiotherapy (RAD) or combination chemotherapy (MOPP)[1], is based on the site and extent of disease involvement and the presence (B) or absence (A) of defined systemic symptoms at the time of diagnosis. An international classification system is used to describe the stage of tumor extent as one of four stages: I, II, III, and IV. Localized disease, stages I, II, and IIIA and can effectively be treated by radiotherapy, while progressive disease, IIIB and IV, must be systemically treated by chemotherapy Table 1 lists the relationship between symptomatology, tumor stage, and expected percentage 5 year disease free survival and 5 year survival rates. The problem for the physician who is managing the patient with Hodgkin's disease is one of sufficiently resolving the uncertainty about tumor extent without harming the patient with invasive diagnostic procedures.

Most patients with this disease present with an enlarged lymph node(s) on the neck or upper torso. Only after the node is biopsied and viewed under an expert's microscope can the diagnosis of Hodgkin's disease be made. Before appropriate treatment can be elected, the physician must determine whether the disease is localized or not. Routine evaluation of a patient for disease extent now costs almost $10,000 and exposes the patients to significant risks of morbidity and mortality. In fact, the culmination of the staging process for a typical patient involves an exploratory laparotomy (LAP). During this surgical procedure the

1. Short hand notation for a four drug regimen consisting of nitrogen mustard, oncovin (vincristine), procarbazine, and prednisone.

patient's abdomen is opened, the spleen is removed, a wedge of the liver is resected, and lymph nodes are removed for microscopic examination. If a LAP is not performed, the patient is said to be "clinically staged" as opposed to "pathologically staged". Published literature suggests that 99 in every 1,000 patients die as a direct result of this operation. Although this particular probability could be adjusted for age, general health, and the skill and experience of the surgical team, the mortality and morbidity possibly resulting from this operation are still important considerations.

Several diagnostic procedures are routinely employed at many hospitals that have a uncertain diagnostic value and costs. First, there are routine laboratory and x-ray studies that are available. Additional costly procedures include: bone marrow biopsy (BMBX), liver biopsy (LBX), gallium scan (GAL), and lymphangiogram (LAG). None of these tests is absolutely diagnostic. BMBX and LBX are diagnostic only if positive; they have large false negative rates due to sampling error. GAL and LAG have both false positive and false negative rates, which may depend on the expertise of the reader. While all these procedures require physician time and hospital space, they also expose the patient to discomfort and the risk serious complications. Table 2 lists the percentage false positive and false negative rates and estimates of mortality rates for each procedure.

C) Goals and motivation

Our goal and motivation has been to produce computer systems which aid physicians making health care decisions. While we do not envision the computer replacing the competent physician, our view has been that there is room for improvement in physician performance. Within this view, the computer is an intellectual tool[1] which supports the

decision making capabilities of the physician. Decision support for the physician managing cancer can constitute a variety of tasks for which the computer is ideally suited. First, the computer can be utilized as an efficient information storage and retrieval system. Electronic storage of patient records is vastly superior to the current hand written and human accessed file systems that are used by almost all health care facilities. Not only can written records illegible, but more importantly they are frequently misplaced or lost.

Second, once the information is in machine usable form, various statistical programs can access the data base and perform analyses that would take humans an extraodinarly long time. Typically such programs are used to search for factors with prognostic significance or study the survival rate for a certain cohort of patients. However, most existing tumor registries, which perform the above two tasks, are designed without really taking a step back and asking "What is this information going to be used for? What information is needed in the decision process?"

Third, the results of computer analysis of retrospective data as well as patient specific data will be that fewer tests will be performed in order to select optimal treatment, thereby decreasing mortality, morbidity, and hospital costs. In contrast to the general tendency of advanced medical technology, the computer holds the promise of actually decreasing health care costs.

Forth, the techniques embodied in our computer system provide a framework for evaluating newly developed tests for their relative merit as additions to or replacements for currently used ones. These methodologies are also useful in the design of new protocols and the re-evaluation of existing ones.

Diagnostic Planning

Finally, data collection and evaluation are an integral part of the clinical investigation of patients with malignant diseases. These data are often incomplete, and a comprehensive analysis focuses attention on those critical areas in need of data, and thereby stimulating clinical research.

It is our thesis that much of the information that would be potentially available to the cancer expert is lost by primitive record keeping and the lack of a model of how to use this information if it were available. Because the decision process is so complicated, much of the available information is basically abstracted into impressions. However, evidence suggests human judgment under uncertainty may have certain systematic biases.(2) For instance, we know that physicians tend to over estimate the importance of a positive diagnostic test by not appropriately accounting for the a priori probability of having a given disease in the population being tested.(3) We believe that a more systematic approach to the cancer patient and available information will produce better medicine and reduced costs.

III) A prototypical System

A) Glossary of Terms

Before proceeding with further descriptions and examples of analysis, the abbreviations and terminology should be clarified. Not surprisingly, some terms have different meanings to computer scientists and physicians. To begin with, Hodgkin's disease is a cancer that affects lymph nodes, and these are not to be confused with decision and chance nodes used in decision trees. Other terms and abbreviations are as follows:

Clinical Specific - information collected before considering any costly tests
- AGE - one of four categories 1 to 15, 16 to 30, 30 to 45, and older than 45.
- SEX - male or female
- SYMPTOMS
  - A - a patient with none of the symptoms listed under B.
  - B - a patient with any of the following symptoms: unexplained fever, night sweats, or weight loss of greater than 10% of body weight during the 6 months preceding diagnosis.
- HISTOLOGIC SUBTYPE
  - NS - nodular sclerosis
  - MC - mixed cellularity
  - LP - lymphocyte predominant
  - LD - lymphocyte depleted
- LOCATION OF PRESENTING NODES
  - LEFT NECK - Patient presents with at least one involved node on the left side of the neck (i.e. left submandibular, anterior cervical, posterior cervical, or supraclavicular). Left neck classification is independent of the presence or absence of nodes elsewhere in the body.
  - OTHER THAN LEFT NECK - Any patient who has no left neck nodal involvement.
- SPLEEN SIZE - by physical exam and scan
- LIVER SCAN - abnormal or normal for size and filling defects

Diagnostic Planning

DIAGNOSTIC TESTS

BMBX - percutaneous bone marrow biopsy
LBX - percutaneous liver biopsy
GAL - gallium scan viewed as a test for subdiaphragmatic node involvement
LAG - lymphangiogram viewed as a test for all subdiaphragmatic node involvement
LAP - exploratory laparotomy with splenectomy

STAGE - unless stated otherwise this is the pathologic stage determined by laparotomy. The following three stages exclude patients that present with localized disease below the diaphragm or with involvement at extranodal sites such as the kidney or lung.

I+II - disease localized above the diaphragm.[2]
III - disease above and below the diaphragm but not including bone marrow or liver involvement.
IV - disease involving the bone marrow or liver

TREATMENT OPTIONS

RAD - radiotherapy to all lymph node bearing areas
MOPP - combined chemotherapy

B) An example session

Consider the following case:

> **A 58 year old male presents with enlarged right cervical and right axillary lymph nodes. Biopsy of these nodes revealed mixed cellularity Hodgkin's disease. The patient denies fever or weight loss, but reported a single episode of drenching night sweats one week before presenting at the hospital. Physical examination revealed splenomegaly which was confirmed by spleen scan. The rest of physical exam was unremarkable. On scan the liver was grossly normal, while alkaline phosphatase levels were reported above normal.**

Now physicians responsible for this patient must decide what further tests to perform; if no further evaluation is to be considered, what is the appropriate choice of treatment. The following discussion and analysis represent a plausible consult produced by our existing Hodgkin's disease decision support system (HDDSS). Of particular concern to

2. The Ann Arbor Staging Classification currently in world wide use includes in this grouping the rare occurence of disease localized below the diaphragm.

these physicians is the reported history night sweats. They wonder if the single episode of night sweats was of unrelated viral origin. Upon presenting the above information to the computer system, two initial estimates of tumor extent are produced.

If the patient is A, then P(I+II)=0.08, P(III)=0.85, P(IV)=0.07

If the patient is B, then P(I+II)=0.08, P(III)=0.61, P(IV)=0.31

From this initial information, the physicians can reach two conclusion about the patient. First, they could feel 92% certain that the tumor has spread to the abdomen. Second, if this patient did have night sweats related to his cancer he would be more than 4 times as likely to have liver or bone marrow involvement than if he was classified as an A. Although these estimations of tumor extent are more accurate estimates than any of the physicians could have produced, their problems still remain: Is further evaluation necessary? How should uncertainty about the patients symptomatology effect their future plans? At this point, the computer system inquires about possible diagnostic options being considered and possible therapies. For this patient, the physicians want to consider all five tests that are listed in table 1. They were satisfied with the false positive and false negative rates that are listed in this table, but they suspected that one the mortality rates listed in table 2 was not accurate. Although a 1% mortality rate for LAP in a 58 year old male seemed appropriate, they felt the risk of LAG was higher than 0.1% mortality because this man had a long history of heavy smoking. It was reasoned that because this man probably had decreased pulmonary capacity the risk of mortality was more likely between 0.1% and 0.5%. With this information entered into the computer, analysis of the decision tree for diagnostic evaluation and treatment selection in Hodgkin's disease is almost ready to begin. Before

proceeding, the computer must ascertain what scale of preference did the physicians want as the basis of their judgments. For a first pass analysis they chose to consider decisions in terms of 5 year disease free survival rates. The computer first analyzed the patient assuming he did have symptoms of night sweats. The amount of information that is actually produced by the computer analysis is quite complex. Figures Ia and Ib represent the diagnostic plan produced for this patient. Figure Ia is the complete plan listing for each recommended procedure (underlined) 1) the patient's expected disease free survival based the information that is available before the test is performed, 2) the probabilities of the test being either positive or negative, 3) the probability of tumor stage if the test turns out to be positive or negative, and finally 4) the procedure or treatment recommended after the test is known to be either positive or negative. Figure Ib is a simplified form of this plan in tree form which does not contain either the expected survival or probabilities after test results. However, figure Ib shows the probability branch of the diagnostic plan being used. This is not shown in figure Ia, but can be calculated from the probabilities of each test result in the branch. In both figures Ia and Ib, the structure of the diagnostic plan is more important than the particular numbers. The indentations in the computer generated printout in figure Ia correspond to the various levels of branching shown in the tree in figure Ib.

BMBX, expected survival 40%
P(-BMBX)=0.94, P(+BMBX)=0.06
P(stage|-BMBX)=I+II 0.08 III 0.65 IV 0.27
P(stage|+BMBX)=I+II 0.0 III 0.0 IV 1.0
If BMBX negative, then perform LBX
If BMBX positive, then treat with MOPP

LBX, expected survival 41%
P(-LBX)=0.94, P(+LBX)=0.06
P(stage|-LBX)=I+II 0.09 III 0.69 IV 0.22
P(stage|+LBX)=I+II 0.0 III 0.0 IV 1.0
If LBX negative, then perform GAL
If LBX positive, then treat with MOPP

GAL expected survival 41%
P(-GAL)=0.63, P(+GAL)=0.37
P(stage|-GAL)=I+II 0.13 III 0.69 IV 0.18
P(stage|+GAL)=I+II 0.02 III 0.68 IV 0.29
If GAL positive, then perform LAP
If GAL negative, then perform LAG

LAP, expected survival 42%
P(-LAP)=0.13, P(+LAP)=0.87
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.79 IV 0.21
If LAP negative, treat with RAD
If LAP positive, treat with MOPP

LAG, expected survival 39%
P(-LAG)=0.5, P(+LAG)=0.5
P(stage|-LAG)= I+II 0.04 III 0.74 IV 0.22
P(stage|+LAG)= I+II 0.01 III 0.63 IV 0.36
If LAG negative, then perform LAP
If LAG positive, then treat with MOPP

LAP, expected survival 40%
P(-LAP)=0.04, P(+LAP)=0.96
P(stage|-LAP) I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP) I+II 0.0 III 0.75 IV 0.05
If LAP negative, treat with RAD
If LAP positive, treat with MOPP

Figure 1a.
Analysis of example for disease free survival assuming patient is B

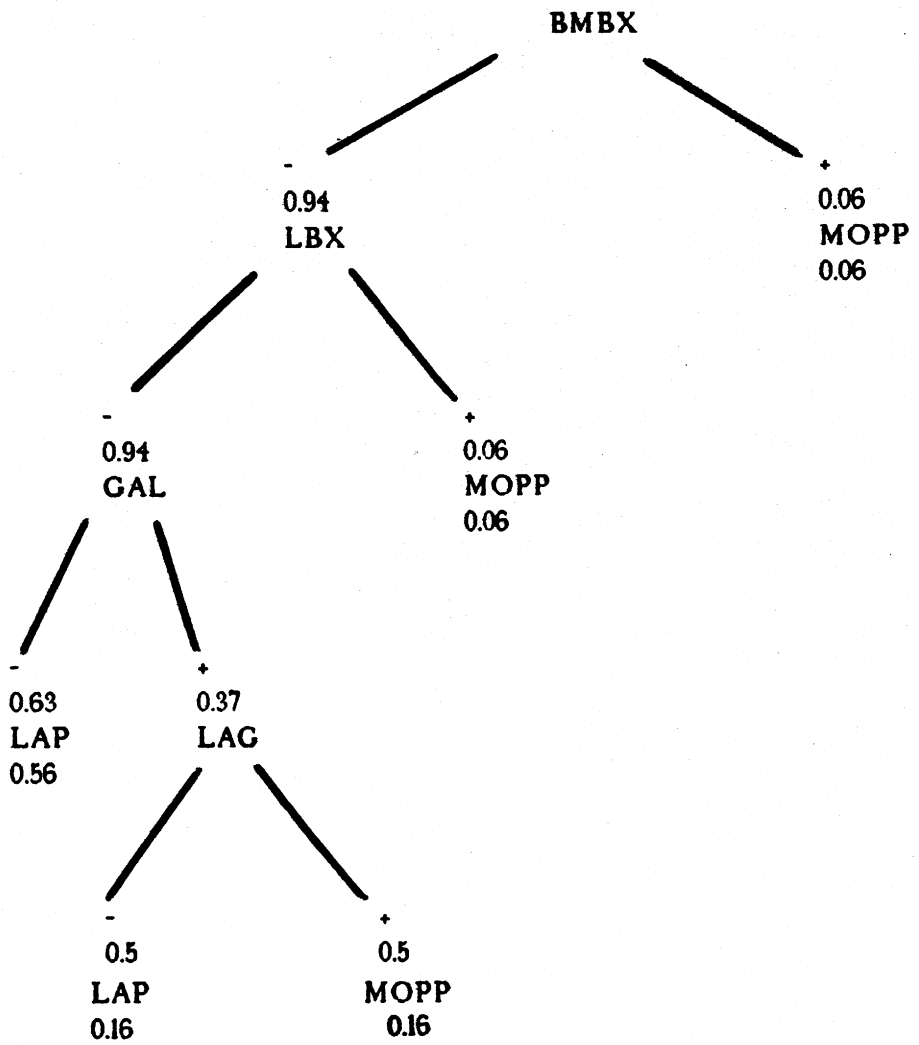


Figure 1b.

Simplification of the disease free analysis for example patient assuming the patient is a B. The numbers directly below a + or - indicate the probability of a positive or negative test result. The numbers at the terminal end of a branch indicate the overall likelihood of a particular branch of the diagnostic plan.

The plan for this patient, based on the assumption that the patient had B symptoms, suggests that he has only a 40% chance of achieving a 5 year disease free period. This survival rate is based on the probability of stage, on the mortality rates incured by using each of the procedures shown in the diagnostic plan in figure 1, and the stage and symptom determined survival rates for each treatment listed in table 3. This plan begins by suggesting a bone marrow biopsy and liver biopsy both of which are only 6% likely to be positive. If both these tests are negative, the probabilities of stage for the patient with symptoms is P(I+II)=0.09, P(III)=0.69, and P(IV)=0.22. Now after two negative testing procedures, the computer has recommended a novel combination of the diagnostic procedures gallium scan (GAL) and lymphangiogram (LAG). If the GAL is negative (p = 0.63), regardless of the LAG's possible results, a LAP would be required to determine appropriate therapy. Therefore, if the GAL is negative, the analysis recommends a LAP without first performing a LAG. However, if the GAL is positive (p = 0.37), perform a LAG. The logic here is that the LAG is about equally likely to be positive or negative. However, if the LAG is positive, the likelihood of localized disease, P(I+II), is 0.4%, so that MOPP could be confidently chosen for this symptomatic patient without a LAP. If the LAG is negative, there remains 4% uncertainty about localized disease, so that LAP is the prudent course of action.

After this first plan was produced, several other analyses were performed to explore how first the absence of symptoms and second a history of heavy smoking would affect the diagnostic plan. We know that if the patient was in fact asymptomatic, the probabilities of stage are: P(I+II)=0.08, P(III)=0.85, and P(IV)=0.07. With these prior probabilities the plan produced by the computer based on disease free survival rates is shown in figure 2.

BMBX, expected survival 63%
P(-BMBX)= 0.99, P(+BMBX)=0.01
P(stage|-BMBX)=I+II 0.08 III 0.85 IV 0.07
P(stage|+BMBX)=I+II 0.0 III 0.0 IV 1.0
If BMBX is negative, then perform a LAP
If BMBX is positive, then treat with MOPP
LAP, expected survival 63%
P(-LAP)=0.08, P(+LAP)=0.92
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.92 IV 0.08
If LAP negative, then treat with RAD
If LAP shows stage III, then treat with RAD
If LAP shows stage IV, then treat with MOPP

Figure 2.
Analysis of example for disease free survival assuming patient is A

So if this patient was in fact asymptomatic, then only a BMBX should be performed prior to laparotomy. All other tests either had a cost that outweighed its benefit in avoiding the inevitable LAP, or the results of the tests would not change the necessity for a LAP.

We can see from the analysis so far that the approach to this patient markedly differs if he is A vs B. Before continuing with analysis exploring the uncertainty about symptomatology, the second uncertain factor, the mortality rate of LAG for a heavy smoker, is analyzed. Since the physicians have stated that the range of mortality rates based on their best estimates for LAG is between 0.1% and 0.5%, a simple sensitivity analysis is performed. Since LAG is not recommended for asymptomatic (A) patients, this analysis is only relevant if the patient is symptomatic. Here the analysis is iteratively rerun for each possible mortality rate between 0.1% and 0.5% incrementing each time by 0.05%. For all mortality rates below 0.4%, the diagnostic plan for this is the same plan shown in figure 1.

Thus diagnostic planning for this patient is insensitive to changes in the mortality rate for LAG that are less than 0.4%. However, when the mortality rate is 0.4% or larger, the diagnostic plan for this patient dramatically changes as shown in figure 3.

BMBX expected survival 40%
P(-BMBX)=0.94, P(+BMBX)=0.06
P(stage|-BMBX)=I+II 0.08 III 0.64 IV 0.29
P(stage|+BMBX)=I+II 0.0 III 0.0 IV 1.0
If BMBX negative, then perform LBX
If BMBX positive, then treat with MOPP
LBX, expected survival 41%
P(-LBX)=0.94, P(+LBX)=0.06
P(stage|-LBX)=I+II 0.09 III 0.69 IV 0.22
P(stage|+LBX)=I+II 0.0 III 0.0 IV 1.0
If LBX negative, then perform LAP
If LBX positive, then treat with MOPP
LAP, expected survival 41%
P(-LAP)=0.08, P(+LAP)=0.92
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.73 IV 0.25
If LAP negative, then treat with RAD
If LAPpositive, then treat with MOPP

Figure 3.

Analysis of the example patient with a mortality rate for LAG greater than or equal to 0.4%.

In the analysis in figure 3 one notices that not only is the LAG not recommended because of its higher mortality rates, but the GAL is also not recommended. For this example patient the usefulness of the GAL and LAG are intertwined. It is the combination of results that may offer a chance to avoid LAP. With the results of the sensitivity analysis

for the mortality rates of LAG, the physicians must decide whether the mortality rate for LAG in the 58 year old male is above or below 0.4%. Although this analysis does not provide a definitive answer for the physician, it does focus complex decisions about the usefulness of LAG and GAL on the mortality rate of just the LAG.

The concern about the patient's night sweats is more complicated to analyze. Not only does symptomatology have an effect on the probability of stage, but its effect on prognosis changes therapeutic strategies. If a patient is IIIA, RAD is the treatment of choice; while MOPP is the treatment of choice for IIIB. Frequently, when a patient with an uncertain finding is managed, physicians will either assume the finding was there or it was not. At this point in the analysis, the computer system tries to ascertain from the physician the probability that the night sweats were not related to the patient's Hodgkin's disease. In this particular case, the several physicians disagree about the nature of the night sweats. Undaunted by their confusion, the computer produces three separate analyses reflecting their differing points of view. Figures 4a, 4b, and 4c are the plans for this patient when P(A) equals 0.5, 0.75 and 0.25 respectively where P(B) = 1 - P(A). Each of these figures first shows the prior probability of stage based on uncertain symptomology and then shows the recommended diagnostic plan.

If P(A)=0.5 and P(B)=0.5, then P(I+II)=0.08, P(III)=0.73, and P(IV)=0.19

BMBX expexted survival 45%
P(-BMBX)=0.99, P(+BMBX)=0.01
P(stage|-BMBX)=I+II 0.08 III 0.74 IV 0.18
P(stage|+BMBX)=I+II 0.0 III 0.0 IV 1.0
If BMBX negative, then perform LBX
If BMBX positive, then treat with MOPP

LBX, expected survival 45%
P(-LBX)=0.96, P(+LBX)=0.04
P(stage|-LBX)=I+II 0.08 III 0.78 IV 0.14
P(stage|+LBX)=I+II 0.0 III 0.0 IV 1.0
If LBX negative, then perform LAP
If LBX positive, then treat with MOPP

LAP, expected survival 45%
P(-LAP)=0.08, P(+LAP)=0.92
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.85 IV 0.15
If LAP negative, then treat with RAD
If LAP shows stage III, then treat with RAD
If LAP shows stage IV, then treat with MOPP

Figure 4a.

Analysis of example patient if physician thinks it is equally likely that the patient is A or B.

If P(A)=0.75 and P(B)=0.25, then P(I+II)=0.08, P(III)=0.79, and P(IV)=0.13

BMBX, expected survival 53%
P(-BMBX)=0.99, P(+BMBX)=0.01
P(stage|-BMBX)=I+II 0.08 III 0.80 IV 0.12
P(stage|+BMBX)=I+II 0.0 III 0.0 IV 1.0
If BMBX negative, then perform GAL
If BMBX positive, then treat with MOPP

GAL, expected survival 53%
P(-GAL)=0.6, P(+GAL)=0.4
P(stage|-GAL)=I+II 0.12 III 0.77 IV 0.11
P(stage|+GAL)=I+II 0.02 III 0.83 IV 0.15
If GAL negative, then perform LAP
If GAL positive, then perform LBX

LAP, expected survival 55%
P(-LAP)=0.12, P(+LAP)=0.88
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.87 IV 0.13
If LAP negative, then treat with RAD
If LAP shows stage III, then treat with RAD
If LAP shows stage IV, then treat with MOPP

LBX, expected survival 51%
P(-LBX)=0.96, P(+LBX)=0.04
P(stage|-LBX)=I+II 0.02 III 0.87 IV 0.11
P(stage|+LBX)=I+II 0.0 III 0.0 IV 1.0
If LBX negative, then perform LAP
If LBX positive, then treat with MOPP

LAP, expected survival 52%
P(-LAP)=0.02, P(+LAP)=0.98
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.88 IV 0.12
If LAP negative, then treat with RAD
If LAP shows stage III, then treat with RAD
If LAP shows stage IV, then treat with MOPP

Figure 4b.

Analysis of example patient if the physician thinks the patient is more likely to be asymptomatic (A).

If P(A)=0.25 and P(B)=0.75, then P(I+II)=0.08, P(III)=0.67, and P(IV)=0.25

BMBX, expected survival 38%
P(-BMBX)=0.98, P(+BMBX)=0.02
P(stage|-BMBX)=I+II 0.08 III 0.68 IV 0.24
P(stage|+BMBX)=I+II 0.0 III 0.0 IV 1.0
If BMBX negative, then perform GAL
If BMBX positive, then treat with MOPP
GAL, expected survival 38%
P(-GAL)=0.59, P(+GAL)=0.41
P(stage|-GAL)=I+II 0.12 III 0.67 IV 0.21
P(stage|+GAL)=I+II 0.02 III 0.70 IV 0.28
If GAL negative, then perform LBX
If GAL positive, then perform LAG
LBX, expected survival 39%
P(-LBX)=0.95, P(+LAP)=0.05
P(stage|-LBX)=I+II 0.13 III 0.70 IV 0.17
P(stage|+LBX)=I+II 0.0 III 0.0 IV 1.0
If LBX negative, then perform LAP
If LBX positive, then treat with MOPP
LAP, expected survival 40%
P(-LAP)=0.13, P(+LAP)=0.87
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.81 IV 0.19
If LAP negative, then treat with RAD
If LAP positive, then treat with MOPP
LAG, expected survival 35%
P(-LAG)=0.44, P(+LAG)=0.66
P(stage|-LAG)=I+II 0.04 III 0.72 IV 0.24
P(stage|+LAG)=I+II 0.01 III 0.68 IV 0.31
If LAG negative, then perform LAP
If LAG positive, then treat with MOPP
LAP, expected survival 37%
P(-LAP)=0.04, P(+LAP)=0.96
P(stage|-LAP)=I+II 1.0 III 0.0 IV 0.0
P(stage|+LAP)=I+II 0.0 III 0.75 IV 0.25
If LAP negative, then treat with RAD
If LAP positive, then treat with MOPP

Figure 4c.

Analysis of example patient if the physician feels that it is more likely to be symptomatic (B).

The first of these plans, 4a analyzes the case where the physicians could not decide at all whether the patient should be classified as A or B, so that P(A)=P(B). Notice that this plan is different than the plan if either the patient was A or B in figures 2 and 1 respectively. Figure 4b is the plan produced for the physician who basically feels the patient is asymptomatic, P(A)=0.75, but the physician can not completely rule out the possibility that the patient may in fact be a B. This plan differs from the previous plan in its interesting use of the GAL. In this plan, the GAL basically determines whether a LBX should be performed before a LAP. Finally figure 4c is for the physician who is basically convinced the patient should be classified as a B, but this physician feels there is a small chance that the night sweats were of an unrelated cause, P(A)=0.25. The diagnostic plan in this last analysis is similar but not identical to the plan for the patient if he is a B, shown in figure 1. The major difference between these two plans is that the latter in figure 4c employs LBX less often.

What has the physician gained after requesting the above kind of consult for planning the staging of the patient with Hodgkin's disease? A lot of numbers and figures have been produced by the computer, but what is the answer? What should be done to the patient? The decisions must still be made by the patient and physician, but the analyses focus the management problems on a few important factors. Although the computer system has no way of judging the patients history of night sweats, it provides a tool by which a doctor can explore the implications of a judgment before choosing a course of action. Each of the plans is designed to maximally utilize each test, so that if a test's results will not change future action, it is not recommended. Furthermore, if the cost of a procedure is

outweighed by its marginal usefulness it also is not recommended. Thus each of the recommended plans requires on the average less diagnostic evaluation to achieve high quality therapeutic results than is currently accepted practice.

However, from the above analyses it is possible to draw some conclusions. First, the more likely it is that the patient's nigh sweats were unrelated to his Hodgkin's disease, the less pre-lap evaluation is recommended. Second, if the mortality rate of LAG is greater than or equal to 0.4% for this patient, a LAG should not be performed. So the complex considerations of staging this patients focus on symptomology and a single mortality rate. Before leaving this example for a detailed description of how such an analysis is produced by a computer system, we should state that the above analysis is by no means complete. All the analyses were based only on a disease free utility criteria. Other criteria such as survival or a combination of survival and disease free survival are available as utilities to drive analyses which produce plans which can differ from the plans based only on disease free survival. In addition, there are many other factors which could potentially stand the scrutiny of a sensitivity analysis as for example the diagnostic accuracy of the various tests.[4]

C) Design and structure of components

The prototypical system in use at the New England Medical Center Hospital consists of four components which were modularly designed in MACLISP[5] for flexible interaction and modification. The physician interacts with one of these components, the Diagnostic Planner which is responsible for the flow of information between the other components: a relational data base, a bayesian estimator of tumor extent and a decision analyzer. This flow of information is shown in figure 5.

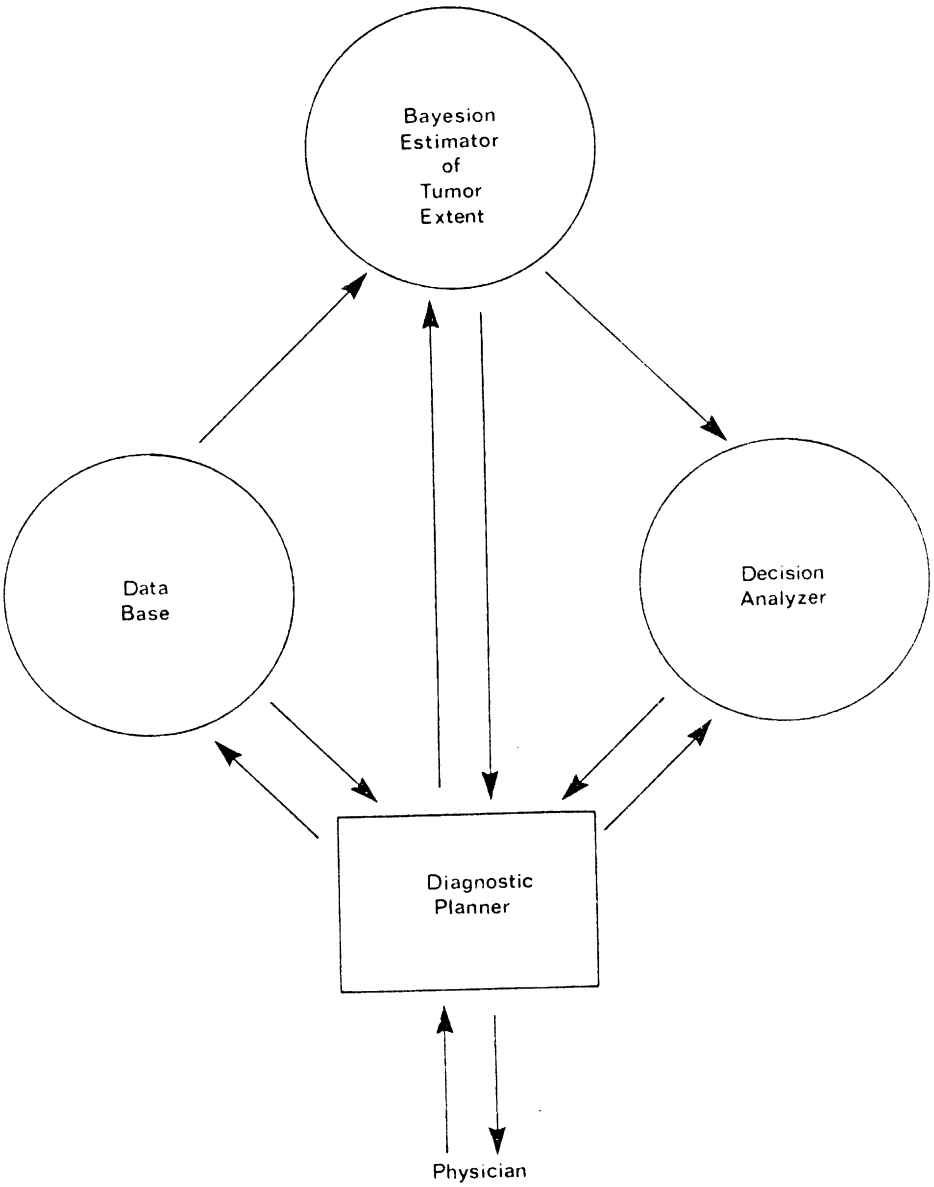


Figure 5.

Flow of information between physician and system components.

1) The data base

Because the data base provides the foundation for the systems quantitative analyses, it is appropriately described first. Our original intention was to provide an interactive library of important data that is needed in the decision process. This data, abstracted from the world literature, could be overriden by the physician-user when local expert judgment differed from the baseline knowledge. In addition to accepting a single value for a particular statistic, the physician might specify a confidence interval. For instance, a mortality rate would be entered as a value between 0.1% and 1%. The uncertainty about the mortality could be evaluated in the following sense: "Does this uncertainty effect diagnostic and therapeutic strategies?" In this way, a physician would test the sensitivity of data estimates in the decision making process.

In time we have expanded our concept of a data base to include actual patient records. We feel that our large collection of patient records increases the credibility of our approach in the medical community and brings our support system closer to being a clinically useful tool. One would ideally like to make decisions based on the most current, accurate information. Some of the medical journals have up to a year delay from date of acceptance to date of publication. Not only could a computerized tumor registry provide more recent information than a medical journal, but it also provides an active rather than passive source of information.

The collection of data directly from the literature involves a number of problems. First each study analyses a small number of patients. Second, the patients are not reported in full detail. Third, results are manipulated and presented according to the desire of the

investigator. Forth, when data from several studies are combined there is always the risk that the data were from different underlying populations of patients. In addition, there is no quaranty that signs, symptoms, laboratory results, diagnostic results, or even histopathological diagnoses mean the same thing from hospital to hospital. In fact, one interesting study has shown that even within a single hospital, experts may disagree about the histological classification of Hodgkin's disease.(6) Not until three or more physicians were forced to agree on a classification before leaving a room, could the investigators demonstrate a significant consistency of cell classification.

Unfortunately, simply having access to a computerized tumor registry does not solve many of the problems inherent in literature data. Tumor registries are all to frequently designed and implemented without proper consideration given to what data could actually be used to make diagnostic and therapeutic plans. Therefore, after the registry is developed it may not contain the data which physicians will want to access. However, in spite of the many difficulties associated with data collection, there is no excuse for wasting information by not electronically storing patient data. With an online tumor registry, data can be analyzed and reanalyzed for any combination of parameters. In addition to providing current best estimates of probabilities which are important in the decision making process, an active tumor registry is a repository for new and accumulating experience. Presumably as the data base grows, the statistical significance of data comparisons also grow.

We began by collecting and abstracting patient records from T-NEMCH and from the world literature. By this method we were able to obtain information on 400 pathologically staged patients. Unfortunately, these patients were reported in varying

amounts of detail without any survival information and were from inhomogeneous populations. The distribution of many of the clinical parameters such as age, sex, and histologic subtype were similar to those distributions that have been reported by others in the literature. This isn't surprising since the data comes primarily from the literature. Since our small collection of data did not contain any information on the prognosis of patients with Hodgkin's, we still sought more complete data. Dr. Henry S. Kaplan, of the Stanford Medical School, had been keeping a computerized file of his patients over the last several years. Although this data was not complete, it did contain prognostic information which could be related to age, sex, histologic subtype, pathologic involvement, and treatment. As previously mentioned, this data base is the best controlled and largest of its kind.

i). Data format

This data which Dr. Kaplan provided consisted of 909 patients strings of information with 11 fields in each string. This positional format was converted into a relational format as shown in figure 6.

377763/25/M/3AS/M-H-S+N-/NS/H5A/AS/1 17 72/-/1 13 75/

transformed into

```
((Stanford-id 377763)
 (age 25)
 (sex male)
 (stage 3AS)
 (symptom A)
 (pathology (bone-marrow -)
            (liver -)
            (spleen +)
            (abdominal-node -))
 (histology nodular-sclerosis)
 (treatment-protocol H5A)
 (Status alive-without-recurrence)
 (First-seen (1 17 72))
 (Relapse)
 (Last-seen (1 13 75)))
```

Figure 6.

Example of a conversion from positional to relational data format.

This patient would be summarized as a 25 year old male with nodular sclerosing Hodgkin's disease. A laparotomy was performed which was only positive for spleen involvement. Total nodal irradiation (Stanford treatment protocol H5A) was administered and the patient has remained in a disease free state for a three year period.

The reasons for transforming the data base into relational form are that the positional format is costly to match into and difficult to add or subtract new features. In the relational format, each patient's clinical parameters are completely cross referenced at the time of entry into the data base and are independent of the order in which they are entered. Thus as this example patient is entered into the data base, an internal lists of all the relations and values are updated to contain a pointer to this new entry. When we want to know the survival of all the males who were treated with a certain drug, the list of all

the males has already been stored. Also a list of all those treated with a certain drug is stored so that a simple intersection provides the appropriate list of patients. Furthermore, if a new relation such as a test result is added at a later date, the computer programs do not need to be modified.

In addition to the patients collected from Stanford and the literature, another 70 patients from the records of the Havard Joint Radiation Center were provided by Dr. Samuel Hellman. This data set contains not only clinical parameters and the results of laparotomy like the data from Stanford, but this data also contains the results of pre-laparotomy evaluation like the results of lymphangiogram. We have also collected 16 records of patients seen at UCLA, which were seen by Dr. Bluming, and 12 records of patients from T-NEMCH. Finally, 131 patients that were not seen at Stanford, but reported in detail in the literature, have been used.[7]

ii). Searching the data base

Matching into this data base is accomplished by evaluating any logical construction. For instance, if one desired to count the number of males with nodular sclerosis Hodgkin's disease who were pathologically IIIA, one only needs to evaluate the expression:

```
(AND (SEX MALE)
     (HISTOLOGY NODULAR-SCLEROSIS)
     (SYMPTOM A)
     (AND (LIVER -)
          (BONE-MARROW -)
          (OR (SPLEEN +)
              (ABDOMINAL-NODE +))))
```

Figure 7.
A logical construction for matching into a relational data base

Diagnostic Planning

This form would retrieve from the data base a list of internal identification numbers of all the patients that met the criteria of the expression. Of course, the physician never see a form like this one. Interactive computer programs ask a series of questions and then these programs construct a form and direct a search into the data base. Frequently, both when inputing data and retrieving it, one would like to have access to varying levels of detail. For example, in figure 6 the treatment protocol is H5A. This treatment is a kind of total nodal irradiation which in turn is a kind of radiotherapy. We have utilized a *Kind Structure*[8] in the data base to help us access different levels of detail as the need arises. Figure 8 depicts the kind structure used for radiotherapy.

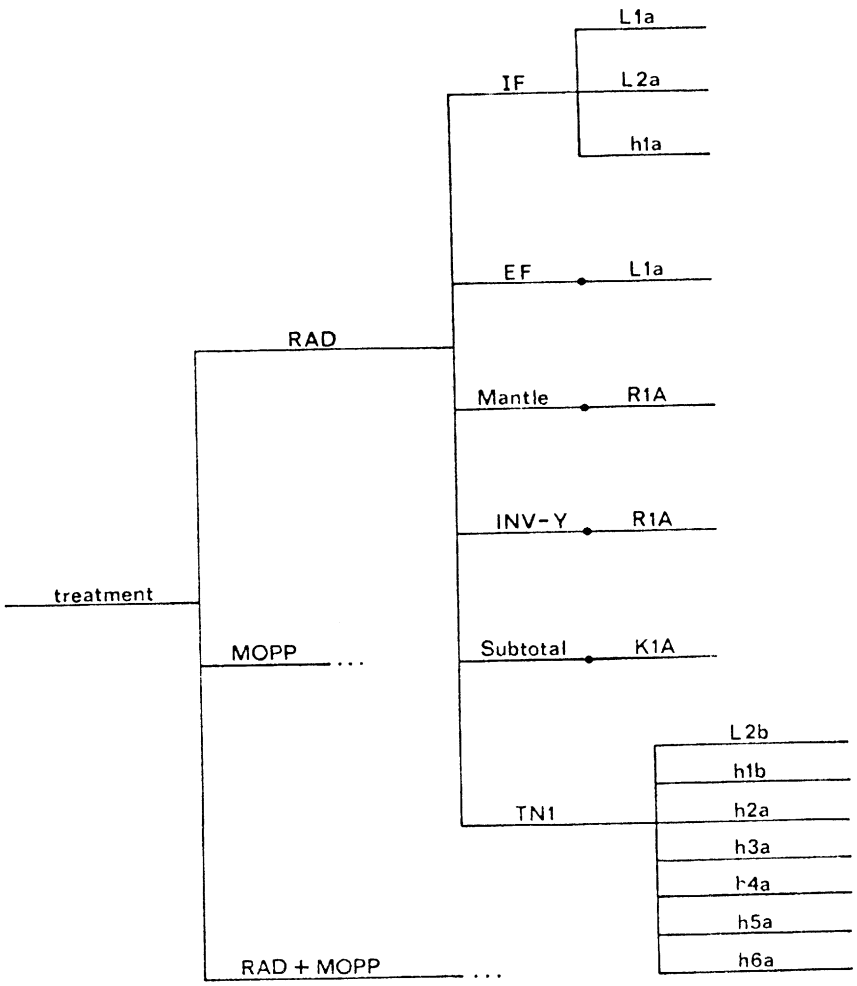


Figure 8.
The kind structure used for radiotherapy

In this figure there are three levels of detail that one could specify about radiotherapy. Before matching a form into the data base, each item is evaluated for its position in the kind tree A terminal value like H5A, is either a concept the matcher can use directly while searching the data base or a logical construction of terms that are terminal values. For any concept that is not a terminal value, a match against this concept would be anything that matched any terminal value below this concept in the tree. So if one wanted to retrieve all patients that were treated with only total nodal irradiation (TNI), we would go one level down in the tree and find 7 terminal values: L2B, H1B, H2A, H3A, H4A, H5A, and H6A. Any patient in one of these protocols would have been treated with total nodal irradiation.

There are several advantages to the scheme of structuring the data base with the kind links between concepts. The data is augmented with a structure that specifies levels of detail for access. This structure is separate from the data base and can be altered without effecting the data. This easily allows one to change the matching strategies as the need arises thus providing a useful facility for interactively computing various statistics. The medical data that can be collected on a given patient is potentially voluminous. The grouping and simplification of the data is obviously necessary. Frequently, as the clinical studies produce a better understanding of the disease process, different groupings of the data may be perferable. In our scheme, all data is stored in detail. The grouping or simplification of data can be designated or redesignated by changing the structure of concepts in the kind tree.

iii). Prognostograms

The results of a search into the data base might either be a calculation of a

conditional probability that is needed to calibrate a model of tumor estimation, or it might be a calculation of survival or disease free survival for all the patients in the data base which matched the request. Standard methods used to calculate life tables or prognostograms[9] produced the disease free survival rates in figure 9 based on the data request in figure 7.

YEARS	TOTAL	LOST	AT RISK	RELAPSE	COND	SURVIVE
1.0	80	15.0	72.5	8.0	0.890	0.890
2.0	57	12.0	51.0	6.0	0.882	0.785
3.0	39	11.0	33.5	1.0	0.970	0.762
4.0	27	9.0	22.5	1.0	0.956	0.728
5.0	17	8.0	13.0	1.0	0.923	0.672
6.0	8	7.0	4.5	0.0	1.0	0.672

Figure 9.
Disease free survival calculations for patients which matched the logical data request in figure 7.

From this table we see that there were 80 patients who matched the request. At the end of a 5 year interval, 17 patients remained in the study and there was a cummulative 5 year disease free survival of 67%. These prognostograms play an important role in the decision making process because it is possible to calculate survival curves for those patients that most nearly match the patient who is being evaluated. In contrast, the survival data that exists in the literature represents a gross grouping of patients which may be quite different from the individual patient being considered. Of course, if the group of patients to retrieved is specified in to much detail, very few patients will be matched and the statistical confidence in the survival rates decreases. However, as the data base grows,

statistical confidence increases and the tailoring of survival rates to the patients under consideration should improve management decisions.

iv). Problems with the data base and data classifications

Since the number of patients which are diagnosed to have Hodgkin's disease each year is about 6000, the number of cases seen at the New England Medical Center Hospitals each year is relatively small. Thus as we mentioned, the data base has been collected from a number of different sources. We have come across several specific problems while collecting a data base for Hodgkin's which should be mentioned. Our most severe problem has been in accessing data in enough detail to properly calibrate our statistical model of tumor spread. Understandably, in order to communicate results, patients must be grouped according to some accepted and uniform criteria. In 1971, a symposium on staging in Hodgkin's disease was held in Ann Arbor which produced a classification scheme which is currently used world wide (see appendix I). Unfortunately this means that all data collected prior to 1972 is reported by a different scheme of classification. The new Ann Arbor classification, while an improvement over the old Rye system, still has some deficiencies which have hampered our continuing investigation. First of all, the grouping was designed to place patients with similar prognoses in the same categories. While this is an important index for a classification scheme, there are other important factors that should be considered when forming a subgroup. In particular the current classification system does not identify some factors which may be important in diagnostic decision making as opposed to therapeutic decision making. Appendix I takes a closer look at the specifics of the classification system.

Another problem which is perhaps a "catch 22" of our study is that we depend on data from laparotomy staged patients. This is because the results of laparotomy are believed to be the "true" stage of the patient. However if our methods are successful, many fewer laparotomies will be performed in the future hence limiting our future data collection efforts. Also during the past several years, laparotomies have been in and out of vogue. In the late 60's only a few treatment centers performed this operation. When they demonstrated that their patients had better survival rates, everyone surgically explored every patient. As emotion gave way to reason, laparotomies were selectively performed on certain classes of patients. Today physicians still argue about when a laparotomy should be performed. The point of this discussion as it relates to data collection is that the data we have received may represent a biased sample of those patients for which a laparotomy was performed. In other words is there something special about those patients who did not have a laparotomy?

2) Bayesian Estimation of Tumor Extent

As mentioned in section II.B diagnostic procedures are used to determine the extent of tumor spread so that appropriate treatment can be selected. Before analyzing the value in a given patient of a diagnostic procedure or therapeutic modality, the physician must first estimate how far the tumor has spread. Although some physicians may be expert at tumor estimation by virtue of vast experience, most physicians are probably much less accurate than the few experts. This must be particularly true in light of the fact that experience with Hodgkin's disease is limited by its relatively rare occurrence. In general, people are poor at estimating probabilities and even worse at combining probabilistic

estimates. Tversky and Kahneman[2] have shown that several of the heuristics that people use to estimate probabilities lead to systematic errors. One of these types of errors "the failure to account for prior information" was dramatically shown to occur in a large teaching hospital at all levels of medical expertise. Schwartz and Gorry[3] posed the following problem to 290 subjects: Suppose a cancer test had a false positive and false negative rate of only 5%. Further suppose that only 5 in 1000 patients actually has cancer. What is the likelihood of cancer given a positive test for a randomly selected patient? More than half of the physician thought the probability was greater than 50%. The actual answer is only 9%. This example clearly shown a failure to integrate prior probabilities of cancer in the population to the interpretation of test results.

i). Bayes' theorem and medical diagnosis

Gorry[10] and many others have shown that Bayes' theorem can be used to revise prior estimates of probabilities given new diagnostic information. Bayes' theorem is a simple formula which is frequently used in probability theory. In the simplest case, suppose a patient either is in a disease state, D, or is not in the disease state, ~D, and an available diagnostic procedure can either be positive, T^+, or negative, T^-. Then Bayes' theorem allows us to calculate the probability of being in disease state D if the test is positive or negative, $P(D|T^+)$ and $P(D|T^-)$ respectively.

(1) $$P(D|T^+) = \frac{P(D)\,P(T^+|D)}{P(T^+)}$$

and

(2) $$P(D|T^-) = \frac{P(D)\,P(T^-|D)}{P(T^-)}.$$

In these equations P(D) is the a priori probability of disease state D, and $P(T^+)$ and $P(T^-)$ are the likelihood that the test will be positive or negative for a given patient. The likelihood of a test result is expressed as follows:

(3) $$P(T^+) = P(D)P(T^+|D) + P(\sim D)P(T^+|\sim D)$$

and

(4) $$P(T^-) = P(D)P(T^-|D) + P(\sim D)P(T^-|\sim D)$$

where $P(T^+|D)$ is the true positive rate, $P(T^+|\sim D)$ is the false positive rate, $P(T^-|D)$ is the false negative rate, and $P(T^-|\sim D)$ is the true negative rate. Thus by estimating the accuracy of a diagnostic procedure with either false positive and false negative or true positive or true negative rates, and by estimating the a priori likelihood of disease states, one can calculate the diagnostic information of either a positive or negative test.

Bayes' theorem can easily be generalized to include more than two disease states or more than two test findings. These types of generalizations increase the number of formula that need to be considered for a given patient and complicate the expressions for P(test-result). In fact, one does not need to be restricted to just test results, but any diagnostic finding would be appropriate. The only restriction on the number of disease states is that they must be mutually exclusive and collectively exhaustive. Obviously the more disease states and test results that are considered, the greater number of conditional probabilities that have to be gathered. For general medical diagnosis, this is one of the limiting factors of using bayesian techniques. As pointed out by Szolovits and Pauker,[11] the real failure of Bayesian technique may be a failure to force the mutual exclusivity on all possible disease states that occur in medicine. If all combinations of multiple disease states are considered and appropriate conditional probabilities are collected, Bayes' theorem

might solve the diagnostic problem in medicine. Luckily for those physicians who have made their livelihood diagnosis, not only is the number of combinations of possible multiple disease states staggering, but the associated data collection would be all but impossible. However, for restricted problem domains where the number of disease states is relatively small, Bayes' theorem can be used with impunity.

ii). Bayes' theorem for Hodgkin's disease

In the case of Hodgkin's disease, the number of disease states that should be considered relates to the number of tumor sites that are important in the decision process. The Ann Arbor staging classification was a natural place to look for predefined disease stages. Basically the limitations of the data that were available to us determined the number of disease states we could use in Bayes theorem. We recognize three distinct tumor stages, I+II, III, and IV. For our purposes, stage I+II represents localized disease above the diaphragm, stage IV is systemic disease in either the bone marrow or liver, and stage III is disease that is not systemic, but on both sides of the diaphragm. Bayes' theorem is adapted for use in the context of Hodgkin's disease in the following manner:

$$(5) \qquad P(S_X|F) = \frac{P(S_X)P(F|S_X)}{P(S_{I+II})P(F|S_{I+II})+P(S_{III})P(F|S_{III})+P(S_{IV})P(F|S_{IV})}.$$

Where $P(S_X|F)$ is the probability of stage x in the presence of a particular diagnostic finding, F; $P(S_X)$ is the a priori probability of stage x; and $P(F|S_X)$ is the probability of finding F in the population of patients with stage x; S_{I+II}, S_{III} and S_{IV}, respectively.

iii). Sequential estimation of tumor extent

Now this theorem can repetitively be applied to an estimation of tumor extent to revise that estimation as every new piece of diagnostic information becomes available. The initial estimation of tumor extent is called the a priori probability. When this estimation is revised in Bayes' theorem it becomes the posteriori probability. However, for the next application of Bayes' theorem, the old posteriori probabilities become the new a priori probabilities. When Bayes' theorem is used in this manner, a basic assumption of independence of diagnostic information is generally employed. While there is no formal necessity to make this assumption, when it can be made it greatly reduces the amount of data collection. We recognized that the independence assumption might not be valid for some of the findings in Hodgkins' disease so we used a technique of judgmental grouping of data. By this technique, whenever two or more findings are strongly suspected to be dependent, joint conditional probabilities are used instead of individual probabilities.

Estimation of tumor extent begins by ascertaining both the patient's histologic subtype and the presence or absence of systemic symptoms. With these two pieces of information, we select select an a priori stage distribution from table I. These two findings have been grouped together because of a suspected dependence between them. Table I was calculated directly from the 509 pathologically staged Stanford patients. In three of the categories, (A LD), (B LP) and (B LD) there were so few patients that estimates were used. These estimates should not affect the overall behavior of our system, only 1% of the patients would present in one of these categories. After the first a priori stage distribution has been selected, Bayes theorem is sequentially applied to several important clinical parameters such

as age, sex, and the location of presenting nodes. Our preliminary investigations have shown that the assumption of independence is reasonable for these parameters.

iv). Evaluation of test results

After the first 5 clinical parameters have been used to produce an a priori stage distribution, results of diagnostic procedures are considered. Now the probabilistic analysis becomes slightly more complicated. Recall in the two disease case, Bayes' theorem uses the false positive and false negative rates of a testing procedure. Unfortunately, in Hodgkin's disease the procedures are not tests for tumor stage, but rather for a specific site of involvement. Three options are available, 1) collect a body of data which reports all pre-laparotomy test results and calculate P(Stage|test result) and use the above form of Bayes' theorem, or 2) do not use the staging classifications and revise probability of particular sites of tumor involvement, 3) develop a model of tumor spread so that each stage can be decomposed into specific site involvement. With these decompositions, a complicated form of Bayes' theorem can be used. When the data are available, option one is by far the simplest. However, conceptually tests are designed to determine tumor involvement and they only incidentally have a relation to the staging classifications that have been imposed. The second options is clearly the most accurate method, but requires large bodies of data on laparotomized patients with all pre-lap testing also reported. The final option is by default the method we chose to implement. Before detailing the mathematics involved, we mention that when the data can be collected, our intention is to implement a system tumor estimation which relies on only sites of tumor involvement rather than stage.

Consider a diagnostic procedure which can be either positive (T^+) or negative (T^-) for a specific site of tumor involvement. For instance a LAG tests for abdominal lymph nodes, a BMBX is a test for bone marrow involvement, etc. For any patient, the tumor site is either involved (+site) or not (-site). Again S_X is one of the stages where x = I+II, III, or IV.

(6) $$P(S_X|T^+) = \frac{P(S_X \& T^+)}{P(T^+)}$$

where from simple rules of probability theory[12]

(7) $$\begin{aligned} P(S_X \& T^+) &= P(+site \& S_X \& T^+) + P(-site \& S_X \& T^+) \\ &= P(+site)\, P(+site|S_X)\, P(T^+|S_X \& +site) \\ &\quad + P(-site)\, P(-site|S_X)\, P(T^+|S_X \& -site) \end{aligned}$$

and where

(8) $$\begin{aligned} P(T^+) &= P(+site \& T^+) + P(-site \& T^+) \\ &= P(+site)\, P(T^+|+site) + P(-site)\, P(T^+|-site). \end{aligned}$$

Now

(9) $$P(+site) = P(I+II)\, P(+site|I+II) + P(III)\, P(+site|III) + P(IV)\, P(+site|IV)$$

and

(10) $$P(-site) = P(I+II)\, P(-site|I+II) + P(III)\, P(-site|III) + P(IV)\, P(-site|IV).$$

Thus $P(T^+)$ can be rewritten by substituting equations 9 and 10 into equation 8 and noticing that $P(T^+|+site)$ is the true positive rate (TP) of the test, T, where TP=1-FN; and $P(T^+|-site)$ is the false positive rate (FP).

(11) $$\begin{aligned} P(T^+) &= P(I+II)\,[P(+site|I+II)\,(1-FN) + P(-site|I+II)\,FP] \\ &\quad + P(III)\,[P(+site|III)\,(1-FN) + P(-site|III)\,FP] \\ &\quad + P(IV)\,[P(+site|IV)\,(1-FN) + P(-site|IV)\,FP]. \end{aligned}$$

The false positive and false negative rates for each test have been abstracted from

the literature and estimated when the literature was deficient[7] and are listed in table 5. The probabilities P(+site|stage) and P(-site|stage) have been calculated directly from the Stanford data base and are listed in table 6. For reasons of accuracy, these values in table 6 are actually P(-site|stage & symptom) and P(+site|stage & symptom). $P(S_X)$ are the a priori probabilities for each stage, x = I+II, III, and IV. Now we are almost ready to calculate the diagnostic value of a test result except for the one remaining terms $P(T^+|S_X$ & +site) and $P(T^+|S_X$ & -site). Now recall that the staging classification is really an artifact that has been devised to aid physicians in analyzing and communicating results, but they in and of themselves do not effect the results of a test. Now $P(T^+|S_X$ & +site) is really a true positive rate for stage S_X and $P(T^+|S_X$ & -site) is a false positive rate for the stage. We make the assumption that the false positive and negative rates remain constant for each stage so:

(12) $$P(T^+|S_X \text{ \& +site}) = P(T^+|\text{+site})$$

and

(13) $$P(T^+|S_X \text{ \& -site}) = P(T^+|\text{-site})$$

so we can finally rewrite equation 6 as.

(14) $$P(S_X|T^+) = \frac{P(\text{+site})\,P(\text{+site}|S_X)\,(1\text{-FN}) + P(\text{-site})\,P(\text{-site}|S_X)\,FP}{P(T^+)}$$

A similar expression could be derived for $P(S_X|T^-)$ and the diagnostic value of test information can therefore be determined. Now as each new test result is reported, these more complex forms of Bayes' theorem are used to revise the estimates of tumor stage. Again, the use of Bayes' theorem in this sequential manner does many assumptions about the independence of test results. Although this may be a reasonable assumption for many of

the diagnostic procedures used, it does introduce a certain error for tests where dependence is suspected. Again, when data becomes available false positive and false negative rates can be calculated for combinations of tests and the pitfalls of the independence assumption avoided.

3) Decision Analysis methodology

While the previous sections are ordered in relation to their use in analyzing the example in section III.B, the entire motivation for the data base and statistical analysis was that the physicians had to make decisions regarding what future tests, if any, to select; and what treatment would be most appropriate. A theory of decision making[13] which provides a framework for handling uncertainty and a "rational" methodology for weighing the costs and benefits of a decision. The basic method of this technique is to break a complicated decision into many of its component parts. In theory, each part is in turn decomposed until the remaining components are so simple that these subdecisions are easily decidable. Then in a rational fashion the simplier solutions are recomposed to form a solution to the original problem. In recent years, these techniques have gained increasing acceptance in medicine as an entire issue of the New England Journal of Medicine[14] was devoted to the subject. While the techniques are fairly simple, the application to a specific problem may be quite complicated. Basically an analysis proceeds in four steps: 1) structure the problem as a decision tree, 2) assign probabilities associated with chance events, 3) asses the utility of an outcome, 4) apply the simple rules of "rational decision-making".

i). Decision Trees

Diagnostic Planning

Decisions trees are composed of decision nodes, chance nodes and terminals. They represent in chronological order all the options open to the decision maker, the chance events that might occur as a result of a decision, and the outcomes that result from a decision and subsequent chance events. As one might imagine, the full complexity of the therapeutic and diagnostic decisions made for a patient with Hodgkin's disease is staggering. If the scope of the problem to be analyzed is not reasonably bounded, a decision tree becomes a decision bush. The utilities that will be used in the analysis provide the first bounds for the problem. Other bounding of the problems reflect the judgment of decision makers. For some problems, analysis need no computational aids. Howver, in complex problems the computer is an invaluable tool. The basic decision made during the evaluation of a patient with Hodgkin's disease is whether to further evaluate the patient's tumor or elect one of the available treatments. This basic Test vs Treat decision is represented in the decision tree shown in figure 10.

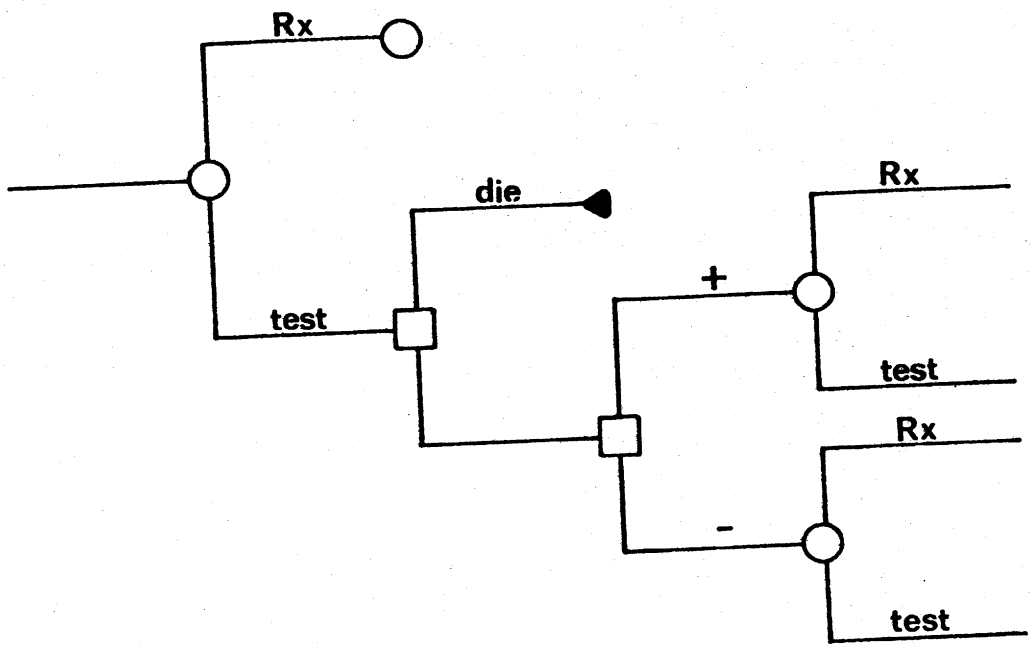


Figure 10.
The test versus treat decision

In this figure, a physician can either select treatment (Rx) or select a test. If the test is selected, a cost is extracted which in our present analysis is represented by the risk of dying from the diagnostic procedure. If the patient survives the procedure, the test can be either positive or negative. However, after the results of the test are known, again the physician must decide whether to further evaluate or to treat. Since the final selection of a therapeutic

modality is driving the diagnostic evaluation process, we began by constructing a decision tree for the therapeutic decision in Hodgkin's disease management. In our first pass at structuring the therapeutic decision, we considered the choice of only two therapeutic modalities, total nodal irradiation (RAD) and combination chemotherapy (MOPP). Figure II represents the therapeutic decision in Hodgkin's disease management. In general, the benefit of a therapy depends on the extent of the tumor. Since we are using a staging classification that corresponds with tumor extent, in figure II, after the decision to elect one of the treatments, there is a chance that the patient would be classified in any one of the three stages I+II, III, IV. Given a particular treatment choice and a tumor stage, there is chance that the patient will relapse within a specified time interval.[3] Then whether or not the patient relapses, there is a chance that the patient will die within that same interval.

3. Because recurrence of tumor is relatively rare after 5 years, the specified interval is usually taken to be 5 years.

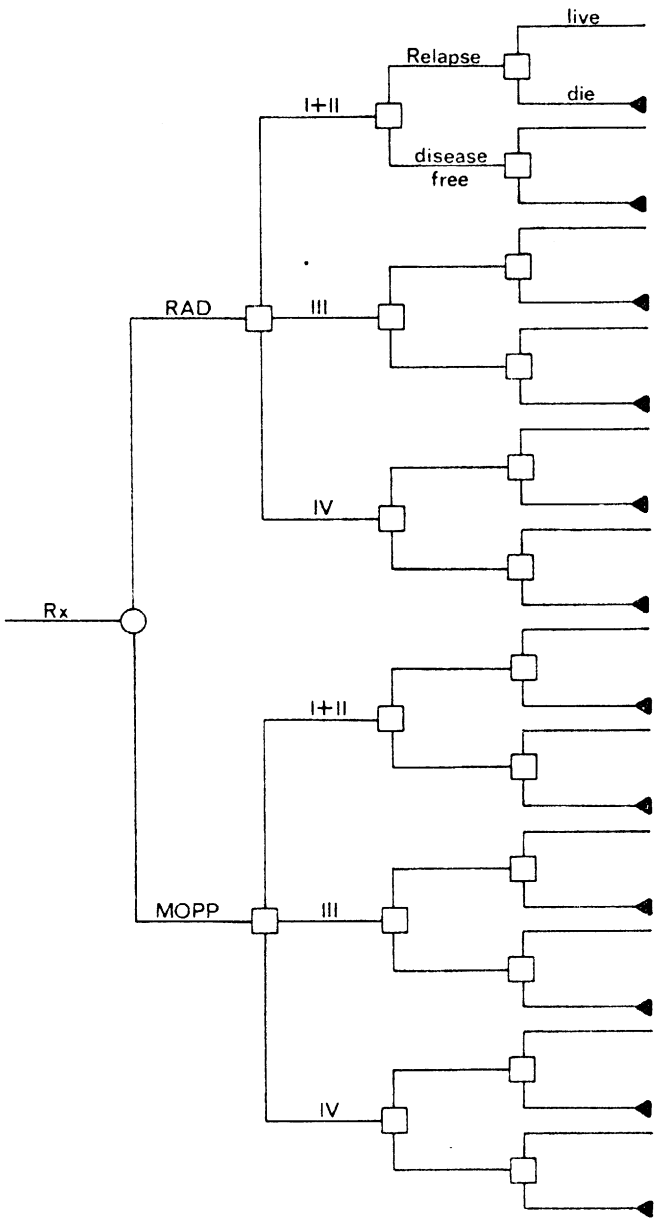


Figure 11.
The therapeutic decision

Diagnostic Planning

Now in order to resolve uncertainty of tumor extent, we consider the choice of any of 5 costly diagnostic procedures: bone marrow biopsy (BMBX), liver biopsy (LBX), gallium scan (GAL), lymphangiogram (LAG) or laparotomy (LAP). If a certain test is chosen, the results are obtained and the physician again faces the *test vs treat* decision with one less test to consider. There is no formal necessity for considering tests only once. In fact, some medical researchers have suggested that doing certain procedures like a BMBX more than once may be beneficial particularly because the large false negative rate associated with BMBX is largely a sampling error. A problem arises however with the independence assumption employed during the sequential use of Bayes' theorem. For this reason tests are considered only once. The tests vs treat decision as tailored to the management of Hodgkin's disease is shown as the decision tree in figure 12.

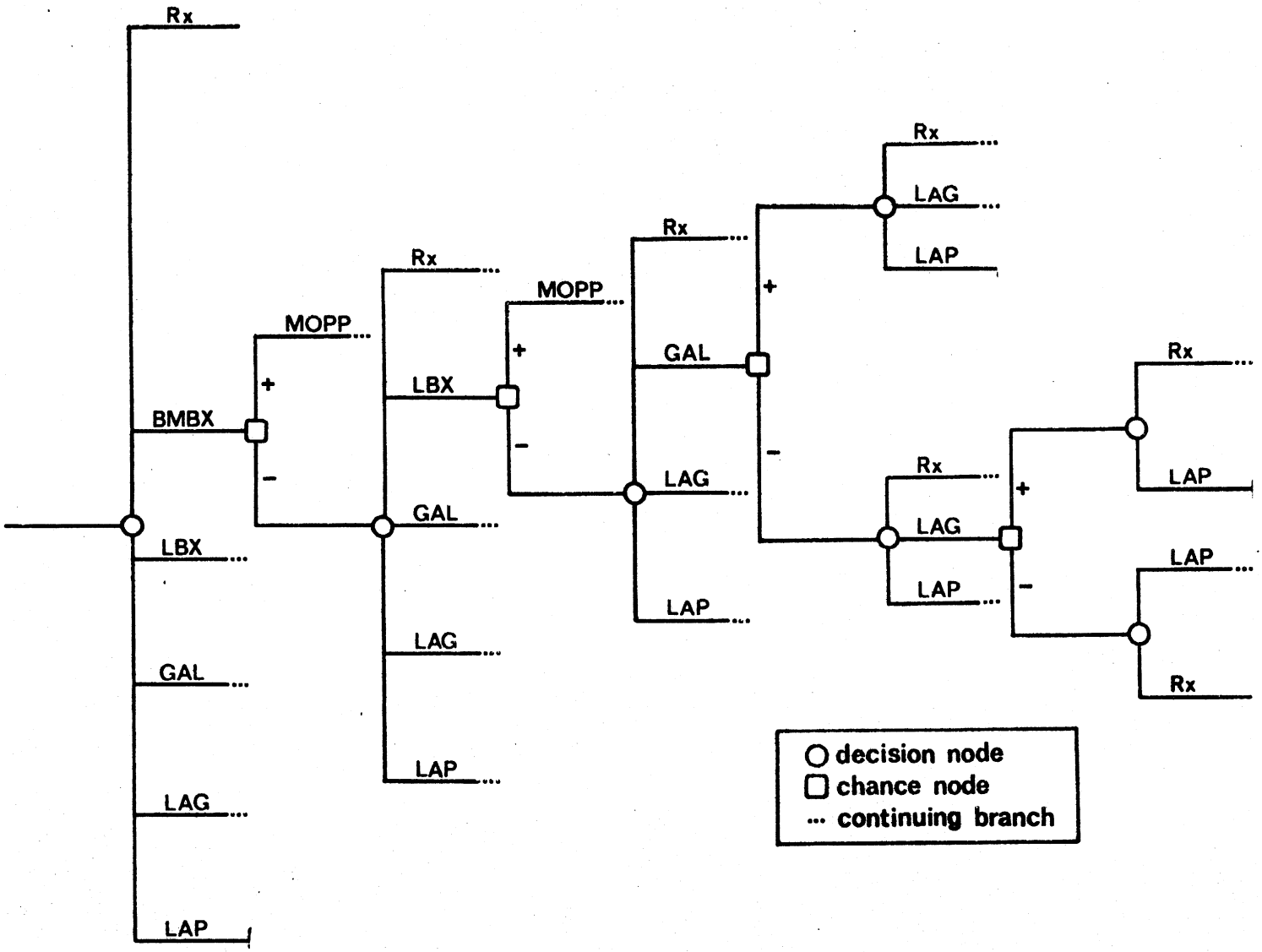


Figure 12.
Decision tree for the diagnostic evaluation of patients with Hodgkin's disease

Because both BMBX LBX have a false positive rate of 0.0 for involvement of bone marrow and liver respectively, a positive test means P(IV) = 1.0 so appropriate therapy

is MOPP. In this tree, the decision to use either LAG and GAL is treated like the test-treat paradigm in figure 10. Now the LAP represents a special kind of diagnostic procedure and is subsequently treated differently than other tests. Because a LAP is both very informative and invasive, it is normally reserved as the last and hence definitive procedure. Because there is really no way of verifying its results, the LAP is usually assumed to be a "perfect test" which would mean that it had a false positive and false negative rate of 0.0. But unlike other tests, this one really samples all the possible sites including the actual removal of the spleen. In figure 13a, LAP is represented as a "perfect" final procedure. If the procedure is negative, the patient is stage I+II and the better of the two treatments, is RAD. If the LAP is positive, then the patient will either be stage III or IV. The appropriate treatment for stage III patients is either RAD or MOPP depending on whether the patient is A or B respectively. If the patient is IV, MOPP is the treatment of choice. While the false positive rate for detection of all types of tumor extent can be confidently assumed to be 0.0, the false negative rate probably is in truth slightly greater than 0.0 for detection of liver, and abdominal nodal involvement. Again this false negative rate is due to sampling error in the surgical procedure. The current system does allow the physician to enter a false negative rate for detecting stage IV disease. When such a value is assigned, uncertainty about tumor stage remains after a negative LAP, so the tree in figure 12 would be modified as shown in figure 13b.

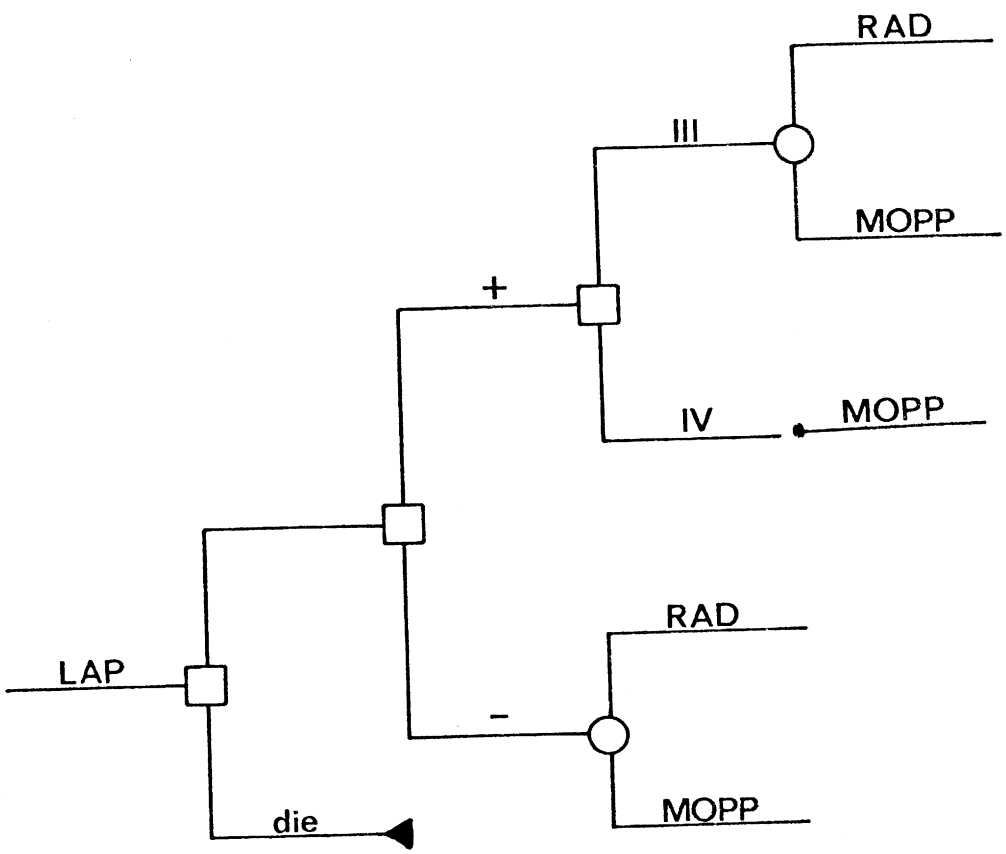


Figure 13a.
Decision tree for LAP when its a perfect test

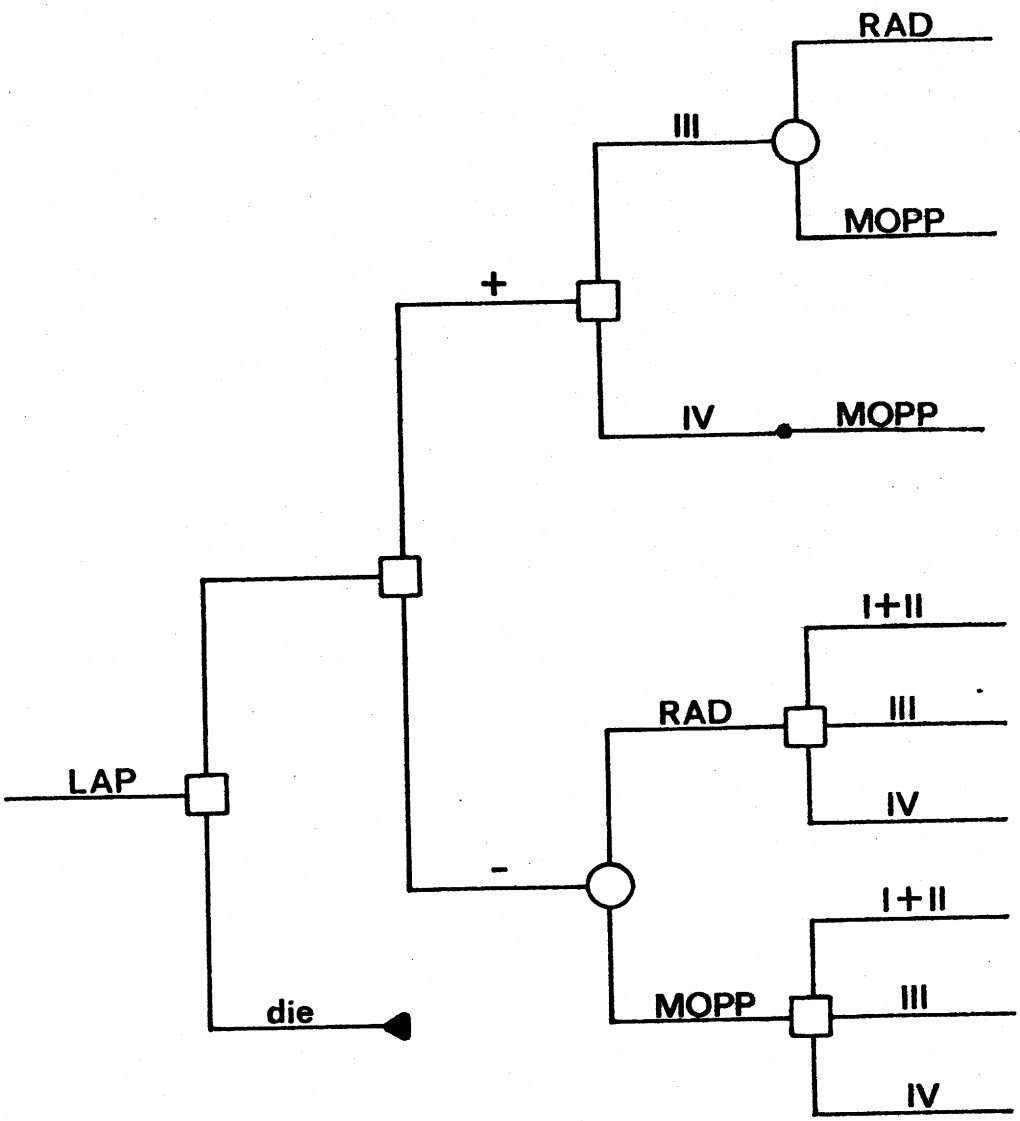


Figure 13b.
Decision tree for LAP when its not a perfect test

ii). Utilities

Perhaps the choice of utilities one chooses is the most important decision made during an analysis. The choice in part determines the structure of the decision tree and consequently what probabilities need to be collected. Furthermore, one decision is chosen over another because of the utilities that are assigned to each. In our analysis, really only two outcomes are currently considered, 1) dying from either a diagnostic procedure or from the disease itself, and 2) relapsing after treatment induced remission. The utility function we use is a weighted average of the probability of surviving 5 years and the probability of remaining disease free during that interval. The utility or expected value of an outcome can be expressed as follows:

$$(15) \quad U(\text{outcome}) = \alpha_1 P(\text{survival}) + \alpha_2 P(\text{disease free survival})$$

where α_1 and α_2 are the relative weights or preferences the decision maker has for basing a decision on survival or disease free survival so that $\alpha_1 + \alpha_2 = 1.0$. Notice that if either α_1 or α_2 is 0.0 then the expected value is simply the expected disease free survival rate or the expected survival rate respectively. The expected survival rates for a particular outcome is first influenced by the mortality rate of any procedure that is to be employed and listed in table 7. The survival rates in the current system are determined by symptom, stage, and treatment as indicated in table 8. However, these survival rates can easily be modified by tailoring these rates with a prognostogram as shown earlier in figure 9. With this type of analysis the survival rates could be adjusted for age, sex, histologic subtype and other interesting prognostic parameters. While the utility function in equation 15 does in

fact fact include some of the "quality of life" factors in considering disease free survival, it basically is deficient in several respects. First it fails to differentiate between dying on the operating table and dying 5 years after the operation and treatment. It is a well established fact that patients are risk adverse to death and greatly perfer dying 5 years hence to dying during an operating procedure. Steve Pauker has investigated this phenomenon while trying to advise his cardiology patients about the risks and benefits of coronary bypass surgery to relieve the pain of angina. His work[15] and the early work of Ginsberg[16] have both indicated that a patients risk adverse behavior can be approximated by an exponential form with an appropriate coefficient to indicate the degree of adversion to a bad consequence. In our scheme, $\alpha_1 P(\text{survival})$ equation 15 would be expanded as follows:

$$(16) \quad \alpha_{1,1} P(\text{survive } 1^{st} \text{ year}) + \ldots + \alpha_{1,5} P(\text{survive } 5^{th} | \text{survive } 4^{th})$$

where $P(\text{survive } i+1^{th} | \text{survive } i)$ is the probability of surviving the $i+1^{th}$ year given that the patient has survived i years already. These conditional survival probabilites are calculated in the prognostogram in figure 9 in the column labled COND. The $\alpha_{1,i}$ are related by the exponential form:

$$(17) \quad \alpha_{1,i} = (1 - e^{-R\ i})$$

where R is a non negative factor indicating the degree of risk adversion.

The utility function in equation 15 does also not account for some of the major morbid complications that affect some therapeutic decision making. Such complications might be a risk of life threatening infection, permanent damage to lungs and heart from

irradiation and the risks of sterility. Two difficulties arise when trying to expand the utility function to include quality of life factors. First, the probabilities of certain complications must be subjectively estimated since no good data exists. Second, relative preferences between outcomes are not necessarily easy to come by. Of course, these preferences differ from patient to patient and as Pauker has pointed out, getting the patient to assign these preferences with the help of a physician may be the real meaning of "informed consent." One simple technique for ascertaining these numbers is through a patient interviewing technique called a lottery.[3,13,15]

For our initial system development we avoided the complex utility questions by using the simpler form of equation 15. Since Hodgkin's disease is fatal if untreated, considerations of survival and disease free survival heavily dominate the decision making process. But while this kind of utility may reflect a large portion of the decision making process, it does not allow us to analyze in depth the special case of children or women of child bearing age. Furthermore, our crude measure of utilities does not allow us to refine our choice of treatment options where morbid factors make a fine difference.

iii). Processing of a decision tree

Once the tree has be structured, the probabilities collected and the utilities assigned, the analysis of the decisions is ready to commence. Two rules ar employed to assign an expected value to every node in the tree. First, any rational decision maker will always choose that option which offers the greatest expected value. Second, the expected value of a chance event is its average expected value. These two rules are called the "folding back" technique by Raiffa[13] and for a given a priori probability of tumor extent, completely

determine what decisions should be made for a given patient given the options in figure 12. For example consider the therapeutic decision in figure 11. For a given set of a priori probabilities of stage and for the appropriate utilities, the expression for the average expected (EV) of RAD and MOPP are written by the second rule as follows:

(18) $EV(RAD) = P(I+II)\, U_{(RAD,I+II,Symptom)} + P(III)\, U_{(RAD,III,Symptom)} + P(IV)\, U_{(RAD,IV,Symptom)}$

(19) $EV(MOPP) = P(I+II)\, U_{(MOPP,I+II,Symptom)} + P(III)\, U_{(MOPP,III,Symptom)} + P(IV)\, U_{(MOPP,IV,Symptom)}$

where (treatment,stage,symptom) is an outcome and U(outcome) is calculated by equation 15 with appropriate probabilities from table 8. When all values are assigned in these two equations, the treatment option offering the greatest expected value would be the treatment of choice if no additional testing were to be considered.

iv). Thresholds

When the expected value of each option is equal, no preference can be asserted between the treatments. If equations 18 and 19 are set equal to each other, then the resulting equation is a therapeutic threshold line in n-space, where n is the number of independent parameters in the analysis. Solving for therapeutic thresholds in one dimension was demonstrated by Pauker[17] and in two dimensions by Safran.[4] Equations for the expected value of diagnostic procedures can also be determined by the "folding back" technique. These equations are considerable more complicated because many more branches have to be considered, and the complicated form of Bayes' theorem, equation 14 must also be used. Never the less, diagnostic thresholds can also be determined for any set of utilities

and probabilities. Numerical analysis techniques and the computer can be used to solve for these more complicated threshold lives[4].

v). Formation of a Diagnostic plan

When a patient first enters our system, the previously defined Bayesian tumor estimation techniques produce and a priori probabilities of stage. These probabilities then determine the analysis of the decision tree in figure 12. At each level in this tree, a diagnostic procedure or treatment is selected. When a diagnostic procedure is chosen, Bayes' theorem is used to revise the a priori probabilities and the process continues. The complete recording of the optimal decisions at every level, their expected value, the results of Bayesian estimation at that level, and the decisions to be made if a procedure is positive or negative constitute a *diagnostic plan* as shown in the earlier example in figures 1 through 4. The result of a single pass through the decision model with a particular set of a priori probabilities produces not only the optimal diagnostic plan, but it also stores for future inspection several *near optimal* plan. Frequently, the choice between one option is based on small differences in their expected values. Since our current utility structure does not reflect any of the morbid consequences in Hodgkin's disease that do not relate to relapse, the comparison of the optimal plan to near optimal ones can provide some valuable information.

vi). Sensitivity analysis

As we have shown, these plans depend not only on a particular a priori value, but also on many other probabilities and utilities that are used in the system. Some of these

numbers have undergone intense scrutiny and we feel represent the best available information. On the other hand, many may not agree with our opinions. Sometimes, findings and test results may not be clear cut. In the early example, the patients night sweats were of questionable origin. In other cases, pathologists may not agree about the proper histologic classification. Finally, the radiologists may report test results that are "highly suggestive" rather than conclusive. Furthermore, a patient or doctor could be uncertain about the weights associated with utilities. The important question to ask about any uncertain piece of information is "**Does this uncertainty change the diagnostic or therapeutic plan in any way?**" With a computerized decision model, specific values can be altered and the analysis rerun. The changing of parameters and the rerunning of an analysis is called a sensitivity analysis and no analysis is really complete without one. The example in section III.B had two sensitivity analyses, one for the uncertainty about the patient's night sweats and a second concerning the mortality rate of a LAG for a 58 year old heavy smoker. While these analyses did not solve the problems of uncertainty, it focus the problems on certain thresholds. If the physicians believed the probability to be above a certain value, then then proceed one way, if not proceed another. It remains up to the physician to decide what he or she believes the actual values to be.

4) Computer implementation

The current system currently operates on Mit's Mathlab Digital Equipment Corporation PDP-10 computer and is written in MACLISP[5] MIT's dialect of LISP. As mentioned in earlier sections there are a number of distinct modules which have specific tasks. The overriding consideration in the design of the system has been to make both the

decision analysis and data base technology accessible to the physician who has no prior exposure to the computer. This means that output is as english like as possible with use of display terminals for plotting and split screen display as a memory aid. Furthermore, input uses command completion to minimize typing and spelling errors of complex medical terminology. In addition several special control characters attempt to salvage analysis in case of a software failure. Besides the many human engineering features, several other implementational features will be mentioned. Details of the data base and matching routines have already been discussed. This section deals with the specific decision analysis algorithms that have be used, where information about the patient is stored, and how the physician is directed from one module to another.

i). Dynamic programming and recursive control structure

No decision trees are stored in our programs. For each analysis the computer starts with the basic test versus treat paradigm. After ascertaining from the physician a few key facts about the patient the computer may alter anyone of several parameters that are important in the analysis. The programs then use the recursive control structure of LISP to generate the appropriate tree structure, that is tailored to the particular patient being evaluated. Furthermore, this same control structure is used to evaluate the tree structure as it is generated. The advantages of this type of scheme is that it allows for a very simple programming definition of decision analysis. A LISP like definition of of the expected survival of some available "tests" given a set of "prior" probabilities could be written as follows.

```
(defun expected-survival (tests prior)
      (maximum
       (mapcar
        (function
         (lambda (test)
                 (expected-survival-1 test prior
                                      (delete test tests))))
        tests)))

(define expected-survival-1 (test prior other-tests)
      (plus (times (probability (positive test))
                   (expected-survival other-tests
                                      (bayes-theorem (positive test)
                                                     prior)))
            (times (times (probability (negative test))
                          (expected-survival other-tests
                                      (bayes-theorem (negative test)
                                                     prior))))))
```

Figure 14.
Recursive definition of expected survival

One can appreciate the relative simplicity of this function which is closely related to the definition actually used by our system. Of course, in order to save "near optimal" diagnostic plans and produce nice print out, more code is needed. However, the decision tree is elaborated and evaluated by the recursive control structure of LISP.

ii). Building a patient model

During the course of getting information about the patient the computer builds a patient model[18] which consists of all the entered information and a limited number of expectations. These expectations are triggered[19] by combinations of items entered about the patient. These expectations are currently used only to suggest interesting sensitivity

analyses. For instance, entering (sex female) and (age 15-to-30) would trigger the expectation that the patient was of child-bearing-age. In addition to storing specific patient information, and the very few expectations, the patient model also store all plans that are to be considered for the patient. It keeps track of those plans that are most likely to be acceptable alternatives to the optimal diagnostic that is first shown to the physician.

iii). Control of information flow

For any given patient, many different analyses could be performed, each one being informative. Unfortunately, one really has to be an expert in decision analysis to get the most of the information that is generated. For the expert user, the programs allow that user to direct the computation in any way that suits him or her. For those that are being initiated, the module called the Diagnostic planner leads the novice though a Bayesian estimation of patient probabilities to a complete decision analysis. This module directs the initial physician interview based on a simple stored branch and flow algorithm. As mentioned, during this interview, a patient model is constructed that serves as a focus for future analysis. A first request for analysis is sent to the "decision analyzer" with the additional stipulation that any plan with comparable expected survival be saved for future reference. After the first pass analysis is through, it looks into the patient model to see if near optimal have been produced. Next the planner looks to see if any of the expectations compel a further sensitivity analysis. Failing to find and imperative, it hopes the physician's interest has not wained and suggest a menu of interesting possibilities such as producing a prognostogram, changing the assumed diagnostic accuracy of some particular testing procedures, changing particular mortality rates, or changing the utility structure by either

considering disease free survival, long term survival or a combination of the two. The purpose of the diagnostic planner is then to provide an interface for the physician which provides access to a variety of computational programs which hopefully help analyze a particular patient.

IV) Results

A) Educational benefits

When we first introduced our system into the hospital a year ago, we had an interesting experience. Recall our earlier discussion of the state of the art cancer management in section II. Here we stated the default decision making mode was to routinely perform 4 or 5 diagnostic procedures culminating in surgical exploration. In fact there even existed a protocol (or decision flow diagram) which required these procedures in the best interests of the patient. At this same time, some of the physicians were having second thoughts about the protocol, but had little more than subjective impressions to base their feelings on. When decision analysis techniques were first presented to the Lymphoma Unit,[4] the general reaction was mild apathy to outright distrust. Within several months we period began to notice a marked variability in terms of the diagnostic plans this group was suggesting. In some instances these plans began to resemble the kinds of plans that our computer system produced. A year after the system was first introduced into the hospital, one heard the comment that the programs don't really do that much for the physician,

4. A group of specialists at T-NEMCH that manage Hodgkin's and non-Hodgkin's lymphoma and meets regularly to discuss difficult diagnostic or therapeutic decisions

"after all, its recommendations are just common sense." We do not mean to imply by this statement that system has gained credibility, to the contrary this comment is used as an excuse not to use the computer programs. However, decision analysis is a paradigm of thought and expression. Unquestionably the physicians that have been exposed to the techniques presented in this paper, take away something that does not require a computer to use. In some sense, physicians that have been exposed to our programs get a feel for the diagnostic potential of a testing procedure, and a feel for diagnostic lookahead. Without the use of our system they more consciously ask questions like "will the results of this procedure change my future decisions?"

Another benefit for training physicians that exposure to an interactive data base really points out, is that in Hodgkin's disease there really isn't a lot of hard data to base ones decision upon. Perhaps this is all the more reason to explore the implications of subjectively assessing data on the decision making process. For many of the educational benefits that our system has provided, it acts more as a catalyst rather than an reactant. In other words, once the student has been exposed to the techniques of decision analysis and seem the basic impact of patient information on the estimation of tumor extent, the techniques can informally be used with out the computer. We do not mean to imply that the computer has been of little value in the study of Hodgkin's disease decision making, but rather there is pedagogic value in what we have done with or without the computer.

One other side effect that we have observed as a result of our work is the general tendency of the physician who use our system to be more precise about uncertainty. Medicine has its own language which has evolved under the guise of facilitating precise

communication. However, look in any medical textbook and you will be amazed at how impresicely they communicate the relations of signs and symptoms to disease. "Usually this happens, but sometimes that..." What does the term "usually" mean, 50% to 60%, 60% to 70% or 70% to 80%? We now observe at Lymphoma Unit meetings doctors demanding probabilistic estimates instead of fuzzy terms of uncertainty. If a test is equivocal, the expert who read the test is pinned down on a range of probabilities that represent his or her best estimate of the test being truely positive. We long for the day when such estimates become part of the medical record.

B) Verification of Bayesian techniques

Besides the educational value of our academic studies, we are uncovering many important relationships among testing procedures and in the data. We are now at the point where we are retrospectively studing how our system would perform in a clinical setting. We are specifically exploring how accurately we can predict stage without a laparotomy. With an admittedly small sample of patients and a fairly simple statistical measure, we seem to have demonstrated that with only 5 pieces of clinical information: age, sex, symptoms, histology, and the presence or absence of involved lymph nodes on the left side of the neck, and a negative lymphangiogram we can predict disease localized above the diaphragm with better than 95% accuracy. It should be emphasized that this result is preliminary and based on a very small and possibly biased sample. We have also discovered in some instances a very bad fit between our predicted probabilities of stage and the actual results of laparotomy in a small patient population. We are now in the process of refining our preditive model and at the same time trying to understand why our predicitions failed.

However, we not only want to know how good our techniques of estimation are, but rather does a little error affect the essence of our diagnostic plans? It is of interest to note that the general tendency error is to underestimate the proportion of patients with localized disease. Our program in 75 studied cases recommended plans that would have enabled the physician to select the appropriate treatment with on the average 3 or less diagnostic procedures. Furthermore, in 50 of the cases it was better than 60% likely that a laparotomy could be avoided. Of course, this assumes that I, II and IIIA patients received total nodal irradiation. Never the less, this indicates a large scale but selective reduction in the levels of diagnostic evaluation of patients with Hodgkin's disease would provide at least the same levels of quality health care. Our own speculation which would be almost impossible to clinically evaluate, is that this reduction in testing would increase survival rates and certainly decrease morbidity rates. The dollar impact of this reduced testing is left to the readers imagination.

C) The role of specific tests

Pernaps the original focus of our project could be best summarized by our effort to evaluate the role lymphangiography plays in the staging of Hodgkin's disease. To complicate matters, almost every major treatment center reported a different accuracy in detecting abdominal nodes. Medical opinion seem to be grossly divided into three categories: 1) do the test, 2) don't do the test, and 3) sometimes do the test. While we are being simplistic about the medicine involved, it is fair to say the problem of evaluating the role of a diagnostic procedure was a poorly defined problem. Therefore, much of the discussion about the role of a procedure was not to the point. What we demonstrated

concerning lymphangiogram was that certain populations of patients can be identified for which the results of lymphangiogram can help avoid a laparotomy. For all other patients, this test has no value in staging patients.[4] Our results demonstrated that for asymptomatic patients with a high probability of localized disease a negative lymphangiogram indicated that a laparotomy would not change the treatment choice. Also for an exceedingly small population of a asymptomatic who have a very high probability of extranodal involvement, a positive lymphangiogram also indicated that a laparotomy would not change the treatment choice. For patients with symptoms the roles were reversed, so that for a population of patients with a high probability of stage III disease, a positive test obviated a laparotomy; and for a very small population of patients with a high probability of localized disease, a negative test obviated a laparotomy. We further analyzed every reported false positive and negative rate and determined that the structured role of lymphangiography as outline above is unchanged by the variable accuracy. Of course, the less accurate the test, the smaller the role.

In addition to analyzing the role of lymphangiography in relation to laparotomy we have considered how other testing effects the use of this procedure. Experience with our system has indicated that a negative gallium scan for A patients and positive gallium scan for B patients enables the lymphangiogram to take on its special role. When a gallium is not performed in these patients, the lymphangiogram is no longer recommended. Conversely, when a lymphangiogram is contraindicated, the gallium has no real purpose. This latter point is demonstrated in our earlier example. In figure 1 both a gallium scan and lymphangiogram are recommended, but when the mortality rate of the

lymphangiogram is greater than 4%, neither procedure is recommended. This special relationship between LAG and GAL is particularly interesting since most medical centers always do one or the other.

We have also explored the effect of a false negative rate for a laparotomy on a diagnostic plan. Here we mean a false negative rate for the detection of abdominal nodes and stage IV disease. Since the spleen is removed completely we assume a 0.0 false negative for spleen involvement. A limited experience indicates two interesting phenomenon one that is expected by "common sense," and another which is less obvious. When the false negative rate of the laparotomy is increased all tests are suggested more frequently since the LAP is no longer definitive. However, the surprising finding is that chronological order of tests became highly non standard. Gallium scans and Lymphangiograms became the first recommended tests. These tests if positive would immediately determine the need for a laparotomy. If they were negative, bone marrow and liver biopsies would then be performed. The ordering of bone marrow biopsy not as the first procedure has never occured when the false negative rate of laparotomy was assumed to be zero. The reasons for this change are still not apparent.

Finally we investigated how the diagnostic plans differed depending on whether the analysis was based on disease free survival or just 5 year survival. While no general rule is forth coming, on the majority of cases studied these differing rates did not affect the structure of the diagnostic plan. In those cases where there was a difference between the analysis based on disease free survival and just survival, the general tendency for asymptomatic patients was to suggest less evaluation if survival was the planning criteria.

This is because in terms of survival these patients are going to do very well if treated with total nodal irradiation without any evaluation. On the other had, symptomatic patients seem to require more evaluation if survival is the planning criteria because according to the data that we had available IIIB patients have a slightly better survival rate when treated with RAD then MOPP. This is not the case for disease free survival. Thus when survival is the criteria, there is more of a drive to differentiate III from IV.

V) Future directions

Through out this report we have tried to point out problems in our approach and the general direction of future work. Three major areas will serve as a focus for our continuing research in the coming year. First, our concept of utility will have to be expanded to include several distinguished morbid complications. This expansion of utility should be particularly oriented towards consideration of localized radiotherapy and the consideration of a combined modality approach involving chemotherapy and radiotherapy. To adequately expand the utility for these new options, at least the following 4 consequences should be considered: 1) decreased immune response, 2) life threatening infections, 3) 2^{o} tumors, and 4) long term complications to the lungs and heart from radiotherapy. In addition to a more accurate view of treatment options, there are two classes of patients which are particularly hard treatment problems and which represent almost a third of all the patients that get Hodgkin's disease. These two groups are women of child bearing age (25%) and children (7%). In order to expand the utility function to include these groups, the risks of sterility from irradiation and the damage to growth of bone and the

regulation of hormones will have to be assessed. Since these two types of patients and particularly children are the hardest to manage, our techniques should be most helpful to physicians in these instances.

In addition to refining and expanding utilities, the problem of matching a utility structure to a patient still remains. While some advocate loteries, their own experience has highlighted many difficulties in this approach. The patient assigned preferences seemed to change from day to day, and different doctors and interns seem to illicit different preferences from the same patient. Furthermore, some patients and doctors have expressed an unsatisified feeling after a lotery has been conducted. Some of our future effort will be directed towards exploring techniques of debriefing the patient to ascertain preference of outcome.

The system's ability to predict tumor extent as well as survival will remain of central importance. We have already noted that this part of the system is realy the only part for which retrospective clinical studies can demonstrate our ability to provide the physician with useful information. As these techniques become more and more accurate, the need for costly diagnostic evaluation decreases. Our investigations have raised more questions about data and prediction that we have answered. One question that is currently being investigated is the nature of independence of findings and test results. We are also considering the implications of combining data from different medical centers. Is the data compatable, is there a similar population of patients, do the physicians at each center read test results in the same way? Many of these questions are directed towards techniques that would allow the building of larger medical data bases, without sacrificing the validity of the data.

Diagnostic Planning

The past bears witness to many extinct computer systems which attemped to support expert decision making. As designers of yet another system should we feel Sisyphean? Why did those systems fail to become part of the decision making process? Must these sytems communicate in a "natural language", and be capable of explaining pseudo thought processes before these systems will be routinely used? These questions were bantered around at the Second Annual Artificial Intelligence in Medicine (AIM) workshop and several opinions were voiced. The opinion that most closely matches our own philosopy can be stated: "When a physician using a computer system can demonstrate that he or she has a decided advantage over the physician who does not use such a system, then this system will be in great demand." Our aim is to provide such an advantage for the physician managing patients with Hodgkin's disease. In time we hope this advantage will clearly be demonstrated. We will continue to integrate decision analysis methodology, data base technology, and man-machine interfacing so that the physician will have a new "decision making" slide rule to sharpen, focus, and hopefully extend decision making skills.

VI) Tables

Table 1

False positive and False negative rates for diagnostic procedures

TEST	SITE	FP	FN
Spleen Scan	spleen	0.2	0.54
Liver Scan	liver	0.23	0.38
Liver Function	liver	0.46	0.18
Bone Marrow Biopsy	bone marrow	0.0	0.28
Liver Biopsy	liver	0.0	0.78
Gallium Scan	abdominal nodes	0.11	0.49
Lymphangiogram	abdominal nodes	0.07	0.41
Laparotomy	everything	0.0	0.0

Table 2

Mortality percent rates for the diagnostic procedures

PROCEDURE	MORTALITY
Bone marrow biopsy	0.001
Percutaneous liver biopsy	0.017
Gallium Scan	<0.0001
Lymphangiogram	0.11
Laparotomy	0.99

Table 3

Probability of 5 year survival and disease free survival given treatment symptom and stage

A symptoms

Disease free survival

TREATMENT	I+II	III	IV
RAD	0.82	0.64	0.10
MOPP*	0.45	0.40	0.35
Survival			
RAD	0.99	0.85	0.35
MOPP*	0.70	0.65	0.60

B symptoms

Disease free survival

RAD	0.69	0.26	0.05
MOPP*	0.40	0.35	0.30
Survival			
RAD	0.89	0.71	0.05
MOPP*	0.62	0.61	0.60

*Estimated with out regard for stage
small sample size

Table 4.

The probability of stage given symptom and histology

A symptoms

	I+II	III	IV
NS	0.67	0.30	0.03
MC	0.41	0.5	0.09
LP	0.71	0.25	0.04
LD*	0.05	0.55	0.40

B symptoms

	I+II	III	IV
NS	0.46	0.40	0.14
MC	0.21	0.3	0.49
LP*	0.71	0.25	0.04
LD*	0.05	0.55	0.40

* estimated values

Table 5.

Conditional probabilities of sex given stage

	I+II	III	IV
Male	0.51	0.58	0.77
Female	0.49	0.42	0.23

Table 6.

Condition probabilites of age given stage

	I+II	III	IV
1 to 15	0.1	0.14	0.09
16 to 30	0.59	0.50	0.26
31 to 45	0.22	0.21	0.23
46 to 75	0.18	0.15	0.42

Table 7.

Conditional probability of presence or absence of left neck
involvement given stage

	I+II	III	IV
left neck	0.68	0.75	0.73
no left neck	0.32	0.25	0.27

Table 8.

Conditional probability of tumor site given stage & symptom

A symptoms

TUMOR SITE	I+II	III	IV
-abdominal node	1.0	0.36	0.28
+abdominal node	0.0	0.64	0.72
-spleen	1.0	0.17	0.0
+spleen	0.0	0.83	1.0
-abdominal node & +spleen	0.0	0.28	0.28
+abdominal node & -spleen	0.0	0.15	0.0
-liver	1.0	1.0	0.19*
+liver	0.0	0.0	0.81*
-bone marrow	1.0	1.0	0.75
+bone marrow	0.0	0.0	0.25
-liver & +bone marrow	0.0	0.0	0.19*
+liver & -bone marrow	0.0	0.0	0.64*

B symptoms

TUMOR SITE	I+II	III	IV
-abdominal node	1.0	0.16	0.1
+abdominal node	0.0	0.84	0.9
-spleen	1.0	0.13	0.0
+spleen	0.0	0.87	1.0
-abdominal node & +spleen	0.0	0.16	0.1
+abdominal node & -spleen	0.0	0.13	0.0
-liver	1.0	1.0	0.19*
+liver	0.0	0.0	0.81*
-bone marrow	1.0	1.0	0.41
+bone marrow	0.0	0.0	0.59
-liver & +bone marrow	0.0	0.0	0.19*
+liver & -bone marrow	0.0	0.0	0.64*

*calculated with out regard for symptom

VII) Appendix I - Staging Classification

ANN ARBOR STAGING CLASSIFICATION

Clinical Staging (CS)

Stage I: involvement of a single lymph node region (I) or of a single extra-lymphatic organ or site (I_E).

Stage II: involvement of two or more lymph node regions on the same side of the diaphragm (II) or localized involvement of an extra-lymphatic organ site and one or more lymph node regions on the same side of the diaphragm (II_E) or (II_S).

Stage III: involvement of lymph node regions on both sides of the diaphragm (III), which may also be accompanied by localized involvement of an extra-lymphatic organ or site (III_E), or by involvement of the spleen (III_S), or both (III_SE).

Stage IV: diffuse or disseminated involvement of one or more extra-lymphatic organs or tissues with or without associated lymph node involvement (IV).

Each stage is subdivided into A and B categories, B for those with defined general symptoms and A for those without. The B classification is given to those patients with any of the following:

1) unexplained fever above 38°C
2) night sweats
3) unexplained loss of more than 10% of the body weight in the six months prior to admission.

Pathological Staging (PS)

The PS classification is to be subscripted by symbols indicating the tissue sampled and the results of the histopathological examination by + when positive for Hodgkin's disease or - when negative. The abbreviations used are as follows:

N+ or N- for other lymph nodes positive or negative by biopsy

H+ or H- for liver positive or negative by biopsy

S+ or S- for spleens positive or negative by biopsy

M+ or M- for marrow positive or negative by biopsy or smear

The above staging classification is currently in world wide use and represents a standardized format for reporting results. Different types of tumor involvement, (i.e. II_S

and II_E) have been grouped within each stage, I, II, III and IV, because of reasonably similar survival rates. Basically, this classification allows one to be quite detailed in reporting actual sites of tumor and with out considering A or B symptoms or the extranodal classification, there are almost 20 possible subclassifications. Of course, many of the possible subclassifications are stage IV subclassifications and never occur, or occur so rarely, that they hardly warrant special analysis. In fact, physicians usually classify the patients by stage and symptom.

While this classification scheme is helpful in many respects, it is deficient in others and in fact hinders an analysis of the diagnostic staging and treatment selection process. Grouping patients by prognosis assumes that the only important therapeutic factors are survival. Since the treatments themselves have severe morbid complications, some patients may wish to opt for the treatment plan which offers a decreased risk of morbidity at the expense of lessened survival. Furthermore, when planning a diagnostic strategy, the probability that a specific site will be involved determines whether a test will be useful to further explore involvement. When radiotherapists actually planning treatment, it is the exact location of involvement which can alter treatment fields.

Although our computer system is will in the future not consider stage at all, a staging classification is important to analyze studies with limited numbers of patients. A slight modification of the current classification system could eliminate some of the deficiencies in the current scheme. Currently, stage I and II are localized disease above or below the diaphragm. While the prognosis for people with localized disease may be similar regardless of the location relative to the diaphragm, patients with disease below the diaphragm

represent a completely different diagnostic and management problem. Furthermore, the classification of abdominal lymph nodes as either present (N+) or absent (N-) is insufficient for many therapeutic considerations. As mentioned above, if the only upper abdominal lymph nodes are involved, possibly only the upper abdomen need be irradiated for stage III_A patients. This alteration in treatment is particularly important if a patient does not wish to become sterile.

We propose to replace the refine the current staging classification by the addition of one new symbol and the replacing of another with two separate symbols. Currently 4 symbols, N, S, H, and M which can either be positive (+) or negative (+) and three stage notations,[5] I, II, and III are needed to specify the stage of a patient. If a sixth symbol for supradiaphragmatic lymph nodes (SN) were added and subscripts were used to indicate the number of nodes involved, the stage could always be determined from the more detailed information if desired. This modification does not really increase the number of subclassifications, but rather identifies specifically localized disease below the diaphragm. Furthermore, we advocate replacing the notation for abdominal lymph node involvement with the two symbols UN and LN when either upper abdominal or lower abdominal lymph nodes are involved respectively. This modification further identifies an population of patients need special analysis. Until patient data is reported in detail, many important questions about tumor spread and prognosis will remain in a judgmental province.

5. The forth stage IV can always be determined when M or H is positive

VIII) References

1. Martin W A: The Computer as a Mental Tool. Am J Dis Child. 127: 791- 792.

2. Tversky a, Kahneman D: Judgment under uncertainty: heuristics and biases. Science 185:1124-1131.

3. Schwartz W B, Gorry G A, Kassirer J P, et al: Decision Analysis and Clinical Judgment. Am J Med 55:459-472.

4. Safran C, Desforges JF, Tsichlis PN et al: Evaluating the role of Lymphangiography for staging Hodgkins' disease patients with the help of Decision Analysis. Accepted for publication N Eng J Med.

5. Moon D, MACLISP reference manual. MIT, Cambridge, MA 1975.

6. Coppleson LW, Factor RM, Strum SB, Graff PW, Rappaport H: Observer disagreement in the classification and histology of Hodgkin's disease. J Nat Cancer Inst 45:731-740, 1970.

7. Safran C, Tsichlis PN, Bluming AZ, Desforges JF: Diagnostic Planning Using Computer Assisted Decision Making for patients with Hodgkin's disease. Cancer. in press

8. Martin WA, Krumland RB: MAPL, A Language for Describing Models of the World. Automatic Programming Internal Memo #6, Project MAC, MIT 1972.

9. Gross A, Clark V: Survival Distributions: Reliability Applications in the Biomedical Sciences. John Wiley & Sons, New York 1975.

10. Gorry GA Barnett GO: Sequential Diagnosis by Computer. JAMA, 205:849-854, 1968

11. Szolovits P, Pauker S G: Remarks on Scoring. Clinical Decision Making Memo April 1976

12. Drake A: Fundamentals of applied probability theory. McGraw Hill, New York, 1967 PP16-19.

13. Raiffa H: Decision Analysis, Introductory Lectures on choices under uncertainty. Addison-Wesley, Reading, Mass. 1968

14. New England Journal of Medicine Volume 293. July 31, 1975

15. Pauker S: Coronary artery surgery: The use of decision analysis. Ann Intern Med. 85(1):8-18, 1976.

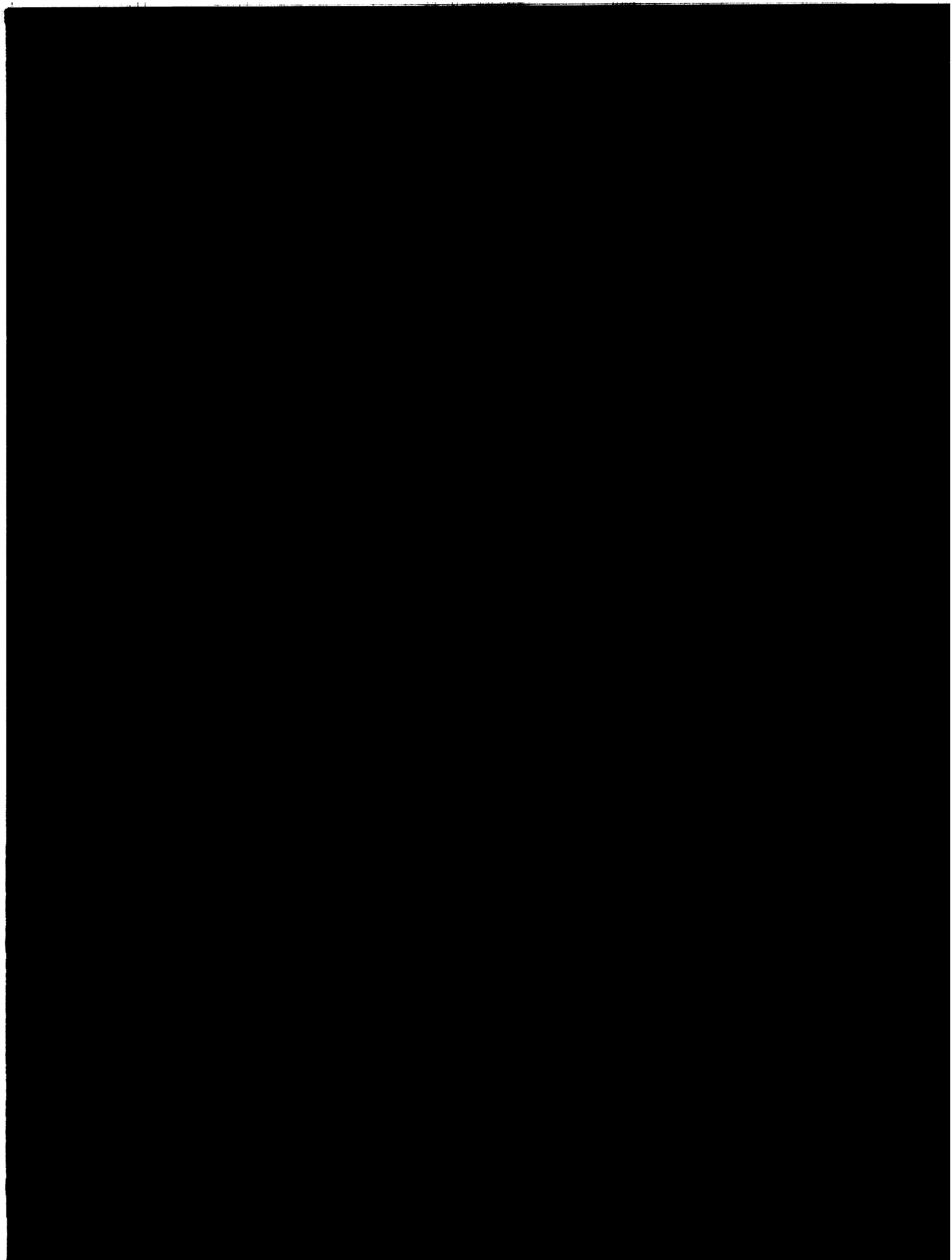
16. Ginsberg A S: Decision Analysis in clinical patient management with an application to the pleural effusion syndrome. Rand Corporation Monograph R-751-RC/LNM. Santa Monica, California, Rand Corporation, 1971

17. Pauker SG, Kassirer JP: Therapeutic decision making: a cost benefit analysis. N Eng J Med 293:229-234, 1975

18. Silverman H: The Digitalis Therapy Advisor. TR-143, Project MAC. MIT.

19. Pauker SG, Gorry GA, Kassirer JP, Schwartz WB: Towards the Simulation of Clinical Cognition: Taking a present illness by computer. Am J med 60: 981-996, 1976.

This blank page was inserted to preserve pagination.



Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

Scanned

Date: 12/5/1995

M.I.T. Libraries
Document Services