1970

Estimating Prognosis with the Aid of a Conversational-Mode Computer Program

Joel F. Rubinstein
Yale University

ESTIMATING PROGNOSIS WITH THE AID OF CONVERSATIONAL-MODE COMPUTER PROGRAM

Joel F. Rubinstein

1970

ESTIMATING PROGNOSIS WITH THE AID OF A CONVERSATIONAL-MODE COMPUTER PROGRAM

Joel F. Rubinstein

B.A. Yale University 1966

A thesis presented to the
Department of Medicine
in partial fulfillment of
the requirements for the
degree of
Doctor of Medicine

Yale University School of Medicine

1970

ACKNOWLEDGMENT

The author would like to thank Dr. Alvan R. Feinstein, whose continued inspiration and guidance have been invaluable in preparing this thesis and have greatly enriched my research experience.

I am grateful to my wife, Nancy Rubinstein, for her candid editorial criticisms.

In addition, the continuing assistance of Carol R. Schimpff, Joyce A. Pritchett, Elizabeth L. Tartagni and Walter A. Ramshaw is gratefully acknowledged.

TABLE OF CONTENTS

I. INTRODUCTION

The decisions of clinical management require an estimate of the outcome of a patient's disease. The background for this prognostication requires experience with many patients, whose case histories are observed over a long period of years and "stored" in the clinician's memory or in written notes.

When performing prognosis, a clinician may find that certain characteristics of his current patient act as reminders of a relatively small group of previous patients. For example, the clinician may recall from his previous experience that elderly men with chronic lung disease and lung cancer did not survive very long. Consequently, if the current patient has these characteristics, the clinician may decide against surgery, but may choose X-ray therapy for symptomatic relief. This traditional approach to the strategy of prognostication has many deficiencies that arise from the limited experience of a single doctor, from the inability of a person to achieve reproducible accuracy in remembering and "retrieving" vast quantities of information, and from the absence of quantification in the results.

The work described in this thesis was done in an effort to improve the scientific state of clinical prognostication. The procedure that has been developed can be used to help a clinician in storing and retrieving his "library" of data about previous patients; in choosing a group of characteristics or properties that adequately describe the current patient with relation

to a segment or resemblance group of the "library" patients; and in determining the quantitative outcome of the resemblance group. To achieve these enumerate analyses of actual clinical cases, the new procedural system was implemented on a digital computer because of the computer's speed, accuracy, and general capacities for storage and calculation of data.

II. COMPUTER METHODS

The system for examining the "library" data of patient records requires two separate interactions with the computer. First, the data on each individual patient must be made available for machine processing, and second, there must be a convenient way to utilize the information stored in the machine.

A. The Library of Data

The library of stored "background" data contains information about the complete clinical course of all the lung cancer patients whose first clinical management occurred at the West Haven Veterans Administration Hospital and the Yale-New Haven Hospital during a particular calendar interval. The library currently contains information on 678 patients for the index years 1953-1959, but future expansions will increase the data base to more than 1000 patients as the 1960-1964 cases are added to the stored collections. Since the computer procedure described here is intended for prognosis, rather than diagnosis, it is based exclusively on data for patients with primary lung cancer.

1. Acquisition of the Data

Before the current project began, the hospital records and other information about the complete clinical course of these patients had been examined, and the data were extracted onto a specially-designed form. The techniques used for obtaining the records and extracting the data are described

in detail elsewhere (1, 2).

2. Codification of the Data

The data on the extraction forms had then been "coded", and entered on machine-readable Hollerith punch cards. The coded information entered on these cards contained a selected subset of data that provided a summary of the entire clinical course for each of the patients. Details of the coding operation are described elsewhere (3). A copy of the Hollerith Coding Form is shown on the next page.

3. Storage of the Data in the Computer

Data for electronic processing can be made available in several ways, including paper cards, tapes, and storage on magnetic discs that resemble large phonograph-record magnetic platters. For our purposes, the data were stored permanently on such discs, for use with the IBM 360/50 computer system at the Yale University Computer Center. The collection of all patient records stored on the disc is called a file or data set.

B. The Time Sharing "Interactive" System

To be satisfactory for clinical work, the proposed computer procedure must fulfill several requirements: (1) it must be easy to use, despite the complexity of manipulating a computer; (2) the clinician must be able to have access to the system near the ward or office where he makes his decisions; and (3) the machinery with which he interacts must be simple and relatively inexpensive. These requirements are necessary in order to avoid the cost of maintaining a supplemental computer technician, and to allow the application of the clinical acumen needed for the important intermediary decisions in manipulating the system.

Because these requirements can be attained by using a computer system,

(Pt. Number)

Cancer of the Lung

Code: Card No. 1: GENERAL SUMMATION

YEAR OF ZERO TIME (1) (2)

AGE (3) (4)

SEX (5)

HOSPITAL STATUS (6)

CUSTOMARY CIGARETTES (7)

OTHER TOBACCO HABITS (8)
Pipes 1
Cigars 2
Snuff or Chewed 4

COMPLAINANT STATUS (9)

IATROTROPIC STIMULUS (10)

CLINICAL GROUP (11)

MORPHOLOGIC GROUP (12)

CO-PULMONARY (13)
Chronic Cough 1
Chronic Dyspnea 2
Active TBC (⊕ AFB) 4

CO-MORBIDITY (14)
Low Respiratory Reserve 1
Significant Deterring Cardiac Disease 2
Poor Condition or Other 4

"BRONCHIAL SX" (15)
Recent Cough 1
Rust, Blood Strks, or Hemoptysis 2
Subjective Wheeze 4

"PARENCHYMAL SX" (16)
Recent Dyspnea 1
Single "Infection" 2
Recurrent "Infection" 4

FEATURES OF INFECTION (17)
Chills, fever or sweats 1
Pneumonia (Clinical Dx) 2
"Cold" or "Flu" 4

RESPONSE OF CHEST "INFECTION" (18)
Chest or "Infectious" Sx better 1
Chest X-ray Better 2

"PARIETAL SX" (19)
Pleuritic Chest Pain 1
Non-Pleuritic Chest Pain 2
UNK re Pleuritic Chest Pain 4

GENERAL (20)
Anorexia 1
Wt. Loss (as complaint) 2
Weakness, Fatigue, or Significant Malaise 4

HPO (21)
Clubbing noted by Pt 1
Signif. Clubbing noted by M.D. 2
Joint Pain due to HPO 4

"MEDIASTINAL SX" (22)
Dysphagia 1
Hoarseness (NMVC or NEVC) 2
"SVC Syndrome" 4

"REGIONAL METASTATIC SX" (23)
Horner's Syndrome 1
Brachial or Peripheral Arm Nerve SSx 2
Pain or Pathol. Fx Sx in Thoracic Bone 4

"DISTANT METASTATIC SX"

(24)
Non-Thoracic Bone Pain 1
Non-Thoracic Pathol. Fx Sx 2
Cerebral Neurologic SSx 4

(25)
Peripheral (not Arm) Neurologic SSx 1
Palpable Surface Metastases 2
Liver or other 4

DOCUMENTED WT. LOSS (26)

PRE-TRANFUSION ANEMIA (27)

SPUTUM CYTOLOGY (28)

GROSS PLEURAL FLUID (29)

MBC CATEGORY (30)

BRONCHOSCOPY GROSS FINDINGS (31)

BRONCHOPATHOLOGY OR LUNG BIOPSY (32)

(33) - (39) Radiographic Interpretations

DESCRIPTION (33)

LOCALIZATION Primary Side (34)

Vertical Location (Field or Lobe) (35)

MEDIASTINAL INVOLVEMENT

Definite Hilar Nodes 1
Definite Mediastinal Nodes 2
Mediastinal "Viscera" 4

☐ (36)

ISO-THORACIC INVOLVEMENT

Pancoast Tumor; or Pleura, Pleural Effusion, Rib, or Chest Wall, same side 1
Other parts of Lung, same side 2
Thoracic Bones 4

☐ (37)

PARA-THORACIC AND CONTRA-THORACIC

Questionable Involvement other side 1
Hilum or Mediast. on other side 2
Any definite structure, other side, or side unspecified 4

☐ (38)

ULTRA-THORACIC INVOLVEMENT

Extra-Thoracic Bone 1
Liver or Abdominal Mass 2
Other 4

☐ (39)

EXTRA-BRONCHIAL PATHOLOGIC EVIDENCE

Para-Thoracic Node Biopsy 1
Pleura: Fluid Cytology or Biopsy 2
Pericardial Fluid Cytology 4

☐ (40)

Thoracic Bone 1
Ultra-Thoracic Site (Not Brain) 2
Brain 4

☐ (41)

PRE-TREATMENT HISTOLOGY

☐☐ (42) (43)

CONTRA-SURGERY REASONS

Juxta-Carinal Location of Tumor 1
Dissemination or Dysfunction of Tumor 2
Poor General Condition 4

☐ (44)

CONTRA-RADIATION REASONS

No Cytologic Proof 1
Cell Type Insensitive 2
Other 4

☐ (45)

FIRST THERAPEUTIC PROCEDURAL DECISION

☐ (46)

PAIRED 1st RX OR REFUSAL

☐ (47)

SURGICAL PROCEDURE AND INTENT

☐ (48)

FURTHER SURGICAL DESCRIPTION

Excision (or Involvement) of other structures 1
Post-op. complix. (other than death) 2
Op. or Post-op. Death 4

☐ (49)

SURGICAL HISTOLOGY

☐☐ (50) (51)

SURGICAL PATHOLOGY

☐ (52)

SUBSEQUENT NEW "CLINICAL" EVIDENCE OF METASTASIS OR RECURRENCE; ATTRIBUTION CERTAIN

Recurrent or Intra-Pulmonary Metastatic Tumor 1
New evid. of Mediastinal Involvement 2
New evid. of Other Involvement 4

☐ (53)

SUBSEQUENT NEW "CLINICAL" EVIDENCE OF METASTASIS OR RECURRENCE; ATTRIBUTION UNCERTAIN

Recurrent or Intra-Pulmonary Metastatic Tumor 1
New evid. of Mediastinal Involvement 2
New evid. of Other Involvement 4

☐ (54)

SUBSEQUENT NEW MORPHOLOGIC EVIDENCE OF METASTASIS OR RECURRENCE

Pericardial or Pleural Cytology of Tissue; Contra-Thoracic Involvement 1
Para-Thoracic Lymph Nodes 2
Chest Wall or Extra-Thoracic 4

☐ (55)

SUBSEQUENT NEW PATHOLOGIC EVIDENCE

☐☐ (56) (57)

SUBSEQUENT TREATMENT

Surgery to Primary Site 1
Radiation to Primary Site or Cytotoxic 2
Surgery or Other Rx to Metastatic Site 4

☐ (58)

MODE OF DEATH

Circumstances Unknown 1
Sudden, or Unexpected 2
Definitely or probably not due to CA Lung 4

☐ (59)

"ASSOCIATED DISEASES"

Co-existent Other Primary Cancer 1
Co-existent Tuberculosis 2
Necropsy Detection 4

☐ (60)

NECROPSY FINDINGS

☐ (61)

NECROPSY TISSUE

☐☐ (62) (63)

SPECIAL COMMENTS

ARF - Note 1
Histology Discrepancy 2
Histology Disparity 4

☐ (64)

PRE-TREATMENT INTERVAL

(* to Rx)

☐☐☐.☐ (65) (66) (67) (68)

SURVIVAL TIME

(Rx to END)

☐☐☐.☐ (69) (70) (71) (72)

☐ (73)

CARD NUMBER [0][1] (74) (75)

☐ (76)

PATIENT NUMBER ☐☐☐☐ (77) (78) (79) (80)

the general operation of such a system will be briefly outlined.

1. An Interactive System

A non-interactive computer system performs its "batch processing" functions without human intervention. The data and the processing instructions are given to the computer simultaneously, and the results are then obtained. By contrast, an interactive system allows human intervention to occur in a "conversational mode" type of activity while the machine is operating. The user can provide new data, choose new options in logic, and give new processing instructions after he sees the results of each stage of the operations. For example, the projected course of a rocket can be computed from data describing the take-off of a space-craft. During flight, however, there is a constant interaction between the computer and the pilots, so that the projected course can be re-calculated as the actual course is observed. As a second example, consider taking a patient's history by machine. A non-interactive system would require answers to be solicited by the operator for all possible details of clinical symptoms. On the other hand, an interactive system would allow a branched selective questioning in depth for symptoms where positive responses are obtained.

In an interactive computer system, communication is transmitted through a terminal, which is a device comparable to a teletype machine. The terminal, connected via a telephone wire to the large central computer, includes a standard typewriter that can be used as the input mechanism for entering information to the computer (by typing in) and as the output device that displays processed information (typed out).

In the prognostic estimation system developed here, the large amount of background data in the "library" can be stored centrally within the main computer, and then called from the terminal.

2. Time Sharing

The type of interactive computation just described becomes economically feasible because of the advanced technology of "time sharing" that enables several terminals to be connected to a central computer while functioning concurrently (4). Although the central memory of the computer is actually "partitioned" for each terminal, the central processing facilities and input-output devices are shared. In the time-sharing operation, the central computer works so quickly that it seems to be under the control of each individual terminal, although each user actually takes turns in access to the operation.

At the onset of the research reported here, this type of time-sharing computer operation was available on CYTOS, the Conversational Yale Terminal Operating System at the Yale Computer Center, using an IBM 360/50 computer. Although the prognostic procedure developed here was designed to operate under CYTOS, a clinician using the procedure need be familiar only with methods used to start the operation of the CYTOS system on the terminal, and to correct errors in typing. A reader desiring further details about the CYTOS system can consult The CYTOS User's Manual (5).

In the prognostic procedure constructed in this project, the interaction between the machine and the clinician is actually controlled by a program, which is a large set of previously written instructions stored along with the data on the disc. The composition of that program, which was one of the main research targets in this thesis, was prepared for the IBM System 360 in the programming language called FORTRAN IV (6). This language was chosen mainly because of its familiarity to the programmer, its widespread availability on many different machines, and the desire for speed in programming rather than maximum efficiency in running time. The prepared

program should be readily adaptable to other time-sharing interactive computer facilities, with or without IBM equipment.

III. ORGANIZATION OF DATA IN THE LIBRARY

The background data in the "library" contain records of individual patients that may be grouped together into specified collections of "charts". The organization of data and formation of "charts" are described in the sections that follow.

A. Properties

The background data for each patient is kept in the form of a "patient record", containing properties and values expressed in codes that can be easily handled by machine.

1. Types of Properties

In the coded form stored in the computer, a patient's record consists of a list of properties that may include demographic features such as age, sex, and smoking habits; clinical features such as physical signs, symptoms, and duration of symptoms; and para-clinical evidence obtained from roentgenography, biopsy, endoscopy, and laboratory tests; and data describing comorbidity, such as the occurrence of severe heart disease in a patient with lung cancer.

2. Values of Properties

Each of the properties for an individual patient is coded in the form of a value that can be expressed in one of four different types of "scale": (1) Existential values are represented by such terms as "present" or "absent" for a symptom, with values of "absent" or "unknown" used for missing data. (2) Nominal values consist of names or other verbal descriptions; examples

of such values are "male" or "female" for the property of sex. (3) Ordinal values are the terms of an arbitrarily graded ranking system, such as the anatomic stages of a cancer. (4) Metric values are represented by numerical counts or measurements, such as a person's age in years.

During the coding of data, each property is expressed in an appropriate value of one of these four types of scales.

B. Formation of a Chart

A clinician creates the "chart" of an arbitrarily-selected new patient by choosing a set of properties and values from the group of properties and values that are available for descriptive purposes. The chart will then be composed of all patients in the background library who fit the stated specifications. The type and number of properties chosen for a particular chart will depend on the purpose for which the clinician intends to use the data. For example, if he wants to know only about survival differences in men versus women, sex would be the only important property; if he wants to know about survival in old men, young men, old women, and young women, the selected properties would be sex and age.

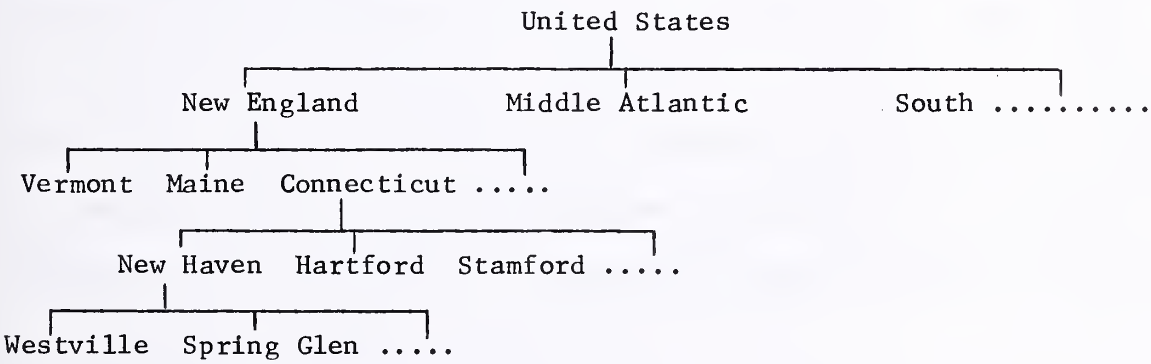
1. The Hierarchical Structure

In order to permit the properties of a "chart" to be expanded into greater detail, or to be condensed into less detail, the data of the "library" have been organized into many hierarchical arrangements.

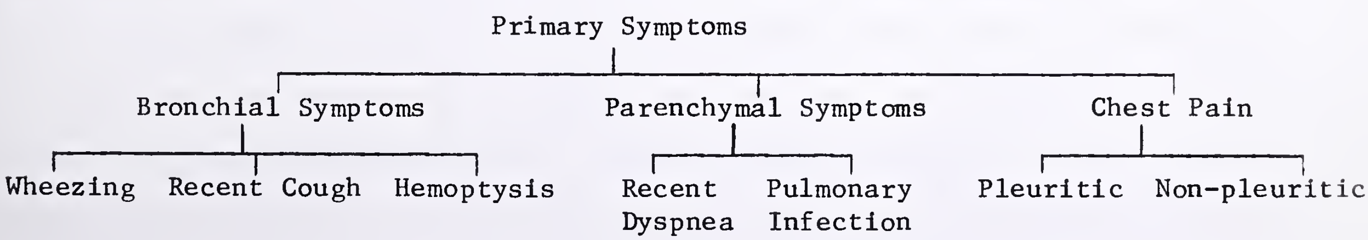
An example of a hierarchical structure can be seen in representative government. The ideas of a city official from New Haven may be significant, but he represents relatively few people nationally. If the man from New Haven is considered representative of Connecticut, he speaks for many more people. His role could be further expanded so that he represents New England,

or the United States. The properties of representing New Haven, Connecticut, New England, or the United States can be called a <u>hierarchy of properties</u>, with representing New Haven the <u>low order property</u> and United States representation the <u>highest order property</u>. One can easily see that there are a number of lower order properties than representing New Haven, including representing Westville, Spring Glen, etc. Alternatively if we start with the highest order property <u>United States</u>, all of its regions, states, cities, and sections of cities will be lower order properties. We have called this organization a <u>tree structure</u> which <u>branches</u> to lower order properties. An example of such a tree is as follows:

<u>Representation Tree</u>



An analogous type of hierarchical tree can be constructed for the symptoms of patients with lung cancer. The classification of such symptoms has been described elsewhere (7), and the symptom tree for <u>primary symptoms</u> is as follows:



In this example, the highest order property in the tree is _primary symptoms_; and _wheezing_, _recent cough_ and _hemoptysis_ are low order properties. Thus, if a patient complains of _wheezing_ he, by definition, is complaining of _primary symptoms_, although a patient with _primary symptoms_ can have many primary complaints other than _wheezing_.

All orders of properties on a tree can appear in a patient's chart, and are not necessarily redundant. If a lower order property of a tree is present, then a higher order property on the same tree is also present and would be redundant in the same chart. However, if the lower order property is absent, the presence of higher order properties would be meaningful. For example, if a patient does not have _bronchial symptoms_ the presence of _primary symptoms_ would imply that he had either _parenchymal symptoms_ and/or _chest pain_.

This type of hierarchical arrangement has been employed for the organization of many of the properties considered in the total data. Among such properties are the arrays of symptoms summarized in the property _clinical stage_, and the array of morphologic evidence summarized in the property _anatomic stage_.

2. Summary Properties

The collected information for a patient can be condensed into a subgroup of hierarchical and non-hierarchical properties that are called _summary properties_. Examples of non-hierarchical summary properties are _age_ and _sex_, which refer to single properties, whereas _clinical stage_ is a hierarchical summary of clinical signs and symptoms. Each summary property may represent one or more of the properties in a chart but there is no overlap in the properties represented. A list of summary properties for a chart presents the highlights of essential parts of a patient's record and permits further exploration when further detail is required. The names of the summary

properties used in this program are: age, sex, smoking, clinical stage, anatomic stage, lateralization, microscopic type, co-morbidity, chronic cough, active tuberculosis, presence of bloody pleural fluid, presence of bronchoscopic mass, contrasurgical indication, and duration of interval of pre-therapeutic symptoms.

C. Operational Tactics in Searching for Resemblance Groups

A clinician who wants to prognosticate for a particular patient must find people in the library population who resemble this patient. As guides in specifying the "resemblance", the clinician will often choose such descriptive features as age, sex, anatomic extensiveness of the tumor, and tumor histology. Even with only these four properties stated, an enormous number of combinations of values becomes possible for different ages, male or female sex, varying degrees of anatomic spread, and diverse histologic types. For example, if age can be specified in four values, sex in two, anatomic spread in five, and histologic type in five, a total of 200 combinations (= 4x2x5x5) of values are possible for these four properties alone. Because a single combination of these values will be possessed by only a few patients, an exact "match" for a particular patient may be hard to find. Thus, although the computer can improve the specificity of recall for a clinician's memory of past experience, the clinician is still left with the problem of deciding how to fit those isolated specificities to the situation of his current patient.

To achieve this goal, a clinician usually engages in a type of reasoning in which he reduces the specificity of description, while maintaining the characteristics that seem to be most important for denoting a "resemblance". In the procedure developed here, the clinician can pursue

such tactics of clinical thought by manipulating the contents of a chart in the search for adequate numbers of patients in the resemblance group. Basically, two patients may be said to resemble each other if their charts have the same values for the same properties. With this concept, the clinician's challenge is to design a chart that includes the important clinical attributes of the patient under consideration, while simultaneously defining a large enough number of patients in the resemblance group of the library. The operations that can be used for manipulating the "charts" are defined and illustrated in the ensuing sections.

1. Range

The tactic of "ranging" values can be illustrated in its application to the property age. A clinician generally does not care about a patient's specific age in years, and usually wants to know only the age in decades, or whether the patient is young, middle-aged, or old. Since the resemblance group for a category of age will be much larger than the number of people at a specific age, the size of the resemblance group can be increased by expressing age in a numerical "range", such as age 45 to 60, instead of using a specific single value for age.

This "ranging" technique is applicable to properties expressed in metric values, such as age and pre-therapeutic interval, but it can also be applied to properties such as clinical stage and anatomic stage that have been expressed in graded ordinal values. For example, suppose the values for anatomical staging have been ranked as: 0, endopulmonic; 1, vicinal; 2, isothoracic; 3, contrathoracic; and 4, ultrathoracic. A patient whose anatomic stage value is 2 (isothoracic) could be cited in the range of one to two (which combines vicinal and isothoracic).

2. Spreading and Merging

The previously described "tree structure" provides hierarchical arrangements for specific lower order properties and for more general higher order properties. The spread and merge operations enable the clinician to move up or down this "tree" in a selected chart. For example, suppose a patient has a recent cough but does not have wheezing and hemoptysis, and suppose each of these three properties is possessed by 100 patients in the library population. For this situation, the higher order property presence of bronchial symptoms might include as many as 300 patients. If the higher order group is considered, the size of the resemblance group for this patient could be tripled, without loss of the implication that the tumor involved the bronchial tree, although the higher order group might include 200 patients who did not have a recent cough. The tactic of disregarding lower order properties and considering a higher order one is called merging.

The tactic of spreading is the reverse of merging. Merging leads to a loss of detail by going "upward" in the "tree", but spreading leads to more details by going "downward". The spread procedure provides greater specificity in a chart by including lower order properties of a property already included. For example, by "spreading" the histologic value undifferentiated tumor, the clinician can define histologic type specifically as undifferentiated small cell carcinoma. Similarly, the "spreading" of the value presence of co-morbid pulmonary disease can lead to the more specific detail, low respiratory reserve. In each of these two instances, the necessary distinctions would be obtained by "spreading" the higher order properties.

3. Dropping and Adding

The resemblance group defined by a chart can be altered by dropping properties that are already present, or adding others that are not contained

in the existing chart. Dropping of properties will usually increase the size of the resemblance group, whereas adding properties will usually decrease it.

4. Unions

A "union" of two properties creates a composite property based on patients who have either the first property, or the second, or both. The union need not be restricted to two properties, and can contain three or more properties. The size of the resemblance group created by a union can be no larger than the sum of people who have each property and value individually. For example, metastatic lung cancer can be diagnosed in several ways. A person can present with symptoms of a pathologic fracture; a biopsy of a cutaneous or subcutaneous lump may reveal carcinoma; or X-rays may show lytic lesions in bone. A group of patients with bony metastases can be constructed as the union of patients who have pathologic fractures and/or X-ray evidence of bony metastases. A union consisting of the presence of the three properties pathologic fractures, positive biopsy, and X-ray evidence of metastases, would create a resemblance group composed of patients with one or more of these three metastatic manifestations. For each patient in the library, the property described in the union is present or absent depending on his values for the individual properties in the union.

IV. PREPARATION OF THE PROGRAM

With this general strategy of operation, a computer program has been written to implement, on an interactive system, the described tactics for prognostication.

A. Outline of the Program

The outline of this program is as follows:

1. The clinician first describes the current patient, either by responding to questions that appear on the terminal, or by entering the patient's data in coded form.

2. After receiving this information, the computer prints out a chart for the patient, consisting of his summary properties and values. Together with each property value, the computer prints out the number of "library" patients who had this value and the associated six-month survival rate for those patients. This table of data is called Chart Number 1 and is the initial "working chart" for the current patient.

3. The computer then tabulates the number of library patients contained in the resemblance group or people who had each of the same property values that appear in the current "working chart".

4. If the clinician believes the size of the resemblance group is large enough, he can ask the computer to indicate the treatments given to these patients, and the associated survival rates for the intervals three months, six months, one year, three years, and five years after treatments. (The clinician would then proceed to step 5.) If the clinician believes the resemblance group is not large enough, he can try to enlarge it by using the tactics described previously. (He would then proceed to step 6.)

5. After survival rates have been obtained, the clinician can either define a new chart and go back to step 3 above, or he can stop the program.

6. After the resemblance group has been enlarged, a new "working chart" exists. The program resumes and continues at step 3.

B. Types of Instructive Message

The program is planned to assist the clinician by displaying messages on the terminal to indicate what the program is currently doing, to give

results of computations, or to offer alternative strategies of operation.

An example of each form of message is as follows:

...PAUSE FOR CALCULATIONS...

This message appears while the central computer tabulates Chart Number 1. The terminal does not appear to be functioning for the few minutes required for the computation, and since the person sitting at the terminal must wait during this period of apparent inactivity, the purpose of the message is to re-assure the user that nothing has gone wrong.

THE NUMBER OF PEOPLE BELONGING TO THE SUBSET OF PEOPLE DEFINED BY CHART NO. 4 IS 3.

This message indicates that the resemblance group defined by the property values in chart number 4 contains 3 people. If the clinician felt that this number was large enough, he could then ask for the survival rates of these patients.

ENTER 1 IF YOU WISH TO AMEND A CHART FURTHER
ENTER 0 IF YOU WISH TO OBTAIN SURVIVAL RATES
>

In this case, if the clinician has defined a resemblance group which is large enough, he can obtain survival rates for the group by typing a "0" after the ">" on the third line, and then hitting the RETURN key. If he types a "1" after the ">", he can continue enlarging the resemblance group. His decision here would depend on previously computed data, and would in turn determine the direction to be taken by the program.

C. Methods of Data Entry

The clinician can enter the data of the current patient either by answering a set of questions on the terminal, or by preparing a coded form and then entering the coded values.

1. Questionnaire Procedure

In the questionnaire procedure, the clinician is "asked" a series of questions by the terminal. Each question is typed individually in full sentences, and requires a response that determines the next question, thus enabling a branching arrangement for collecting appropriate details in

certain types of information.

The illustration in Fig. 1, which is taken from the "print-out" during the operation of the program, shows the questions and responses for a part of the interrogation.

In the internal construction of the computer program, provision is made to check the clinician's answers by searching a list of acceptable answers for each question, and then printing an error message if the cited answer does not match one in the list. For example, the values for sex can be recorded either as 3 (for male) or as 4 (for female); if the user types a 0, 1, or 5, an error message will appear on the terminal. A different type of error would occur if the clinician had indicated that the patient smoked, but had then answered no to each of the questions about smoking cigarettes, cigars, and pipes. In this case, an error would be noted and the original smoking quesion would be asked again. This type of check is necessary so that "higher order" properties such as smoking, and "lower order" properties such as smoking cigars, smoking pipes, or smoking cigarettes, will have consistent values.

2. Coded Input

An alternative method of data input is available to avoid the time needed for answering all the questions. For this option, the patient's data must first be coded on the form "Cancer of the Lung - Code Card No. 1: GENERAL SUMMATION" (see insertion after p. 3). This form, which was used for the original coding of the data contained in the library, contains numbered squares that represent eighty columns of a Hollerith card. After the patient's data have been coded on this form, the program at the terminal asks for entry of the codes in the first 45 columns and in four other columns of the card.

```
HOW OLD IS THE PATIENT IN YEARS?
>65

ENTER SEX OF PATIENT AS 3 FOR MALE OR 4 FOR FEMALE.
>3

HAS THE PATIENT EVER SMOKED ANY FORM OF TOBACCO?
>1

HAS THE PATIENT BEEN A SMOKER OF CIGARS?
>1

HAS THE PATIENT BEEN A SMOKER OF PIPES?
>0

HAS THE PATIENT EVER SMOKED CIGARETTES?
>1

HOW MANY PACKS PER DAY (AVERAGE) DID THE PATIENT SMOKE BEFORE THE APPEARANCE OF THE FIRST SYMPTOMS OF LUNG CANCER?
ENTER NUMBER FOLLOWED BY A DECIMAL POINT.  IF UNKNOWN, ENTER 999.
>1.

DID THE PATIENT STOP SMOKING CIGARETTES BEFORE THE APPEARANCE OF THE FIRST SYMPTOMS OF LUNG CANCER?
>0

DID THE PATIENT CHANGE CIGARETTE SMOKING HABITS WITHOUT ACTUALLY STOPPING BEFORE THE FIRST APPEARANCE OF LUNG CANCER
SYMPTOMS?
>1

HAS THE PATIENT HAD A NEW COUGH OR A CHANGE IN PATTERN OF A CHRONIC COUGH?
>0

HAS THE PATIENT HAD RUSTY SPUTUM, BLOOD STREAKS, OR HEMOPTYSIS?
>0
```

Figure 1

Each question is printed by the "terminal" in capital letters and each response is typed by the clinician on the next line after the "terminal" prints a ">". The character "1" represents yes, and "0" represents no.

In this illustration, the value for the property age is given as 65, and the value for sex is male (coded as 3). The clinician is then asked about smoking. If the patient under consideration had not smoked, the third question would have been answered as "0", and the next question would have involved the property cough. However, in this case yes was the answer to smoking and so the questions then branched to explore specific smoking habits. Also, because the patient smoked cigarettes, the program was arranged to ask questions about the number of packs smoked and changes in smoking habits.

This coded method of entering the patient's characteristics is considerably faster than the questionnaire technique, but does not provide the explanatory instructions of the questionnaire, and requires a knowledge of the coding system. An example of the coded entry is shown in Figure 2 (see next page).

D. Cataloguing of Properties

1. Mnemonics

The information received either from the questionnaire or the coded data is then allocated by the arranged computer program. In this allocation, the information becomes expressed as the values for a list of 137 properties that cover the scope of the patient's characteristics. In order for these properties to be manipulated later at the typewriter terminal, each property must be suitably identified. One method of identification could have been to give each property a number such as 1 for age, 2 for sex, and 25 for chronic cough. This numerical method of identification, however, would require constant reference to a table associating the properties and numbers, because no one could remember all the numbers.

To avoid the nuisance of this procedure, and the concomitant likelihood of errors, we decided to identify the properties with alphabetical abbreviations, called mnemonics, that can be up to twelve letters long, and that are designed to be remembered easily. Some mnemonics are very simple, such as PIPE for the property pipe smoking; and others are more difficult, such as ULTRATHORXR for ultrathoracic X-ray, which implies "radiographic demonstration of tumor outside the thorax". Most of these mnemonics are easy to remember, although a list must be maintained for reference. In actual usage of the mnemonics, mistakes involving the wrong representation for a property have been rare.

```
run cardin

ENTER BELOW DATA FROM THE FORM ENTITLED
CANCER OF THE LUNG-CODE:CARD NO. 1 :GENERAL SUMMATION"
ENTER THE DATA IN THE COLUMNS INDICATED
WITHOUT LEAVING ANY UNNECESSARY BLANKS
AND HIT THE RETURN KEY
COLS(1-8)
>68244113
COLS(9-16)
>11211111
COLS(17-25)
>222777111
COLS(26-35)
>111111111
COLS(36-44)
>11111011
COL 45
>1
COLS(65-68)
>0019
```

Figure 2

An alternative method for describing a new patient. The patient's data are first coded numerically on an external format, called "Card No. 1 ...". The coded numbers are then entered at the terminal in groups of columns as shown here.

2. Storage of Property Values

For storage in the computer, the values for each property have been assigned numerical representations, because the original "library" of data contains numerical values, and because such numbers are convenient for manipulation by the machine. These numerical values can be expressed in any of the four types of "scale" described earlier. Existential values are assigned 1 for present, 0 for absent, and 2 for unknown. Arbitrary numbers are assigned for nominal values, such as 3 for male, and 4 for female. Properties expressed in ordinal values have been assigned a graded series of numbers; for example, clinical staging (with the mnemonic CLINSTGE) has the values 0, 1, 2, and 3, representing the stages of asymptomatic, primary, systemic, and metastatic, respectively. Finally, metric data, such as the property age in years are numerically expressed directly in the required units.

E. Display of Chart Number 1

The first main output of the program, after entry of data describing the current patient, is a mnemonic listing of this patient's values for the summary properties. For convenience, the values are typed in alphabetic form. This tabular list of properties and values is called Chart Number 1, and it defines the initial resemblance group. The basic chart also includes a column that indicates the total number of people in the "library" who had the same value for each summary property. The last column of the basic chart indicates the number of people in each group who survived more than six months, and the six month survival percentage for that group. An example of this print-out is shown in Figure 3 (see next page).

The third column of the example printed in Figure 3 indicates the number of people in the "library" or "base population" who had the same value

...PAUSE FOR CALCULATIONS...

THE CURRENT LIBRARY CONTAINS 678 PATIENTS.

LISTING OF CHART NO. 1

VARIABLE	CURRENT PT.'S VALUE	NO. OF BASE POPULATION WITH SAME VALUE	NO. AND PERCENT OF 6 MONTH SURVIVORS
AGE	65	33	12 (36%)
SEX	MALE	602	246 (41%)
SMOKING	CIGARTTS	608	251 (41%)
CLINSTGE	SYSTEMIC	222	94 (42%)
ANATSTGE	ENDOPULM	201	131 (65%)
LATERALZTN	UNK	2	0 (0%)
MICROTYPE	NONE	211	111 (53%)
COMORBIDITY	PULMONLY	81	32 (40%)
CHRCOUGH	ABSENT	412	163 (40%)
ACTIVETB	ABSENT	660	271 (41%)
BLOODYPLFLD	ABSENT	637	273 (43%)
BRONCHOMASS	ABSENT	536	217 (40%)
CONTRASURG	ABSENT	308	182 (59%)
PRETHERINT	3.0	13	4 (31%)

Figure 3

This example contains all of the summary properties, including demographic variables (age, sex, and smoking); clinical variables (clinical stage and pretherapeutic interval); para-clinical variables (anatomical stage, radiographic lateralization, microscopic type, bloody pleural fluid, and mass at bronchoscopy); comorbidity variables (general comorbidity, chronic cough, active tuberculosis), and presence of surgical contraindications. This list of properties, which appears for any patient entered by the clinician, depends on the universe of properties and not on the specific patient, although the values of each property will be unique for the current patient, as shown in the second column of this example. For further details, see text.

as the current patient. These results are particularly helpful to the clinician in choosing which property values to change for the construction of resemblance groups. In the instance shown here, for example, the presence of the property radiologic lateralization with the value unknown, limits the resemblance group to at most two patients in the library. This property value would therefore have to be altered if the size of the resemblance group is to be increased.

The fourth column of Figure 3 aids the clinician in making crude prognostic estimations based on the best and the worst of the individual property values. In this instance, the property anatomic staging has the value endopulmonic, which means "no evidence of anatomic spread beyond the lungs". In the base population, the 6-month survival rate for patients with an endopulmonic anatomic stage was 65%. On the other hand, the patient's age of sixty-five seems to be a relatively bad prognostic feature, since the 6-month survival rate for this group in the base population was only 36%.

F. The Resemblance Group

After each chart is constructed, the clinician wants to know the size of the resemblance group; i.e. the number of people in the "library" who have the same values for the properties cited in the constructed chart. The computer arrives at this figure by "searching the library file", and comparing the values of each property in the constructed chart with those of each library patient. In this search, the computer ignores any properties that are not cited in the current chart.

For example, after Chart Number 1 appears, the computer automatically calculates the resemblance group for this chart by "matching" the current patient's values, as cited in this chart's properties, with the values of

the summary properties of the "library" patients. Because there are so many different values to be matched, there are seldom any patients in this first resemblance group. In the example portrayed in Figure 3, only two patients in the "library" have unknown lateralization and only 13 patients have a pretherapeutic interval of three months. Because these property values are independent, there is very little likelihood that they would occur simultaneously with all the other values in another person.

For this reason, after noting the absence of any patients in the resemblance group of the first chart, the clinician would then proceed to amend that chart in an effort to increase the size of the resemblance group.

G. Amending a Chart

On the interactive system, the tactics for altering charts are implemented with a set of "commands" that the clinician types on the terminal. Each command is typed together with one or more property mnemonics that indicate the property or properties to be manipulated. After a complete command is typed, the clinician pushes the carriage return button on the terminal, and the desired alteration is automatically made in the chart. The available commands, which were described earlier, include add, drop, merge, range, spread, and union. For example, the user might type "spread primarysx" if he wished more detail of lower order properties for the property primary symptoms. This procedure would add to the chart the properties in the tree illustrated earlier with mnemonics "recent cough", "hemoptysis", and "wheezing" having the values of the current patient.

For convenience in operation, several other commands are also available. These commands include change, which allows the clinician to alter values that may be erroneous; include, which creates a chart consisting of

only the properties listed after the command; list, which gives an interim listing of the chart and its values; and tab, which signals the completed creation of a new chart, and tells the computer to tabulate the resemblance group.

The constructed program contains provision for the computer to make certain checks on the clinician who is altering a chart. For example, during the attempt to range a value for a property, if the clinician cites an interval that does not include the original value, the message "YOUR PATIENT IS OUT OF THE RANGE" is typed. If a mnemonic is spelled incorrectly, the message, "MNEMONICS WRONG" appears. Also, if the user tries to spread a property that does not appear in a tree or is a "lowest order property" on a tree, the machine types "NO SPREAD POSSIBLE". An example of print-out for the amending procedure is shown in Figure 4 (see next page).

After "1" was typed for the list command in Figure 4, an abbreviated listing of the new chart appeared. The new chart is marked No. 2 because it was the second chart to be formed. The properties contained in Chart No. 2 are listed in mnemonic form, and the associated values are in numerical form. The tab signal, given by the character "t" on the last line of the example in Figure 4, indicates that the new chart is completed and that the number of patients in the resemblance group can be calculated.

H. The Estimation Table

Proceeding in this manner from one chart to the next, the clinician will eventually obtain a resemblance group that seems large enough for its therapeutic and survival results to be meaningful. The exact number that will be regarded as "meaningful" cannot be specified according to any statistical pre-conceptions, and will depend on various considerations that enter into

```
***************************
ENTER A COMMAND AFTER A >
>drop lateralztn
>range age(50,75)
   551 PATIENTS HAVE A PROPERTY IN THIS RANGE (6 MONTH SURVIVAL RATE=  42%).

>range pretherint(3,6)
   165 PATIENTS HAVE A PROPERTY IN THIS RANGE (6 MONTH SURVIVAL RATE=  35%).

>l
 LISTING OF CHART NO.  2
   MNEMONIC           VALUE
 SEX                      3
 SMOKING                  2
 CLINSTGE                 2
 ANATSTGE                 0
 MICROTYPE                0
 COMORBIDITY              1
 CHRCOUGH                 0
 ACTIVETB                 0
 BLOODYPLFLD              0
 BRONCHOMASS              0
 CONTRASURG               0
            RANGES
   MNEMONIC          LOW       HIGH
 AGE               50.00      75.00
 PRETHERINT         3.00       6.00
>t
```

Figure 4

In this example, each command is typed after a ">". Furthermore, all of the commands can be abbreviated by typing only the unique first letter as shown by the letter "l" for list, and "t" for tab. The sequence of changes shown here was made in Chart Number One of the preceding example in Figure 3. First, lateralization was dropped from the chart, because its value was unknown and only two people had this value. Next, age and pretherapeutic interval were ranged so that age was between 50 and 75 years, and pretherapeutic duration of symptoms was between three and six months. As noted in this example, whenever a new range is established, the computer automatically tabulates the number and 6-month survival rates of library patients who have property values in that range. In the case of the property age shown in this example, we might consider narrowing the range because too many members of the base population (551 out of 678) fall in the age range of 50 to 75 years.

the deliberations of clinical judgment. For patients with particularly uncommon manifestations, a resemblance group of two or three patients might be "meaningful", whereas patients with relatively common manifestations might require a much larger number. When the clinician believes that he has a meaningful number, and requests the display of the survival percentages, they are computed for the entire group of patients and also for each type of therapy received by those patients, including various combinations of surgery, radiation, and chemotherapy, as well as no anti-neoplastic treatment. An example of the print-out at this stage of the program is shown in Figure 5.

The table shown in Figure 5 contains many interesting items of information. First, it shows the treatments offered to these patients. In this case, chemotherapy was not given to any of the patients, probably because the tumor was anatomically localized. Although chemotherapy is not listed, three of these patients were treated with radiotherapy, and three others were untreated. (The co-existence of major non-neoplastic pulmonary disease may have been the reason for not performing surgery in these patients.) Second, the results show that the entire group of patients had a relatively favorable outcome with a 64% 6-month survival (as opposed to 41% survival at 6 months for lung cancer patients in general), and one person survived five years. Third, the table shows the results of the individual therapies. Thus, surgery for the primary tumor was performed on five patients despite their poor pulmonary status. Three of these patients [60%] survived for at least one year, but one of the five patients died in the post-operative period, presumably as a result of the surgery. Finally, various therapies can be compared. For this small sample it might be inferred that short term survival was improved in quantity and probably in quality if no surgery was attempted, but that long term survival required surgery.

...PAUSE TO COMPUTE SURVIVAL RATES...

ESTIMATION BASED ON CHART NO. 5

TREATMENT GROUPS	TOTAL	3 MON	PERCENTAGE ALIVE AFTER: 6 MON	1 YEAR	3 YEAR	5 YEAR	POST OP DEATH
ALL	11	73%	64%	45%	9%	9%	9%
SURGERY	5	80%	60%	60%	20%	20%	20%
-ALONE	4	75%	75%	75%	25%	25%	25%
-THEN XRAY	1	100%	0%	0%	0%	0%	0%
XRAY	4	75%	50%	25%	0%	0%	0%
-ALONE	3	67%	67%	33%	0%	0%	0%
UNTREATED	3	67%	67%	33%	0%	0%	0%

Figure 5

In this example, survival percentages were computed for 11 patients with property values that are a subset of the Chart Number One shown earlier. These 11 patients had concurrence of: clinical evidence of systemic symptoms, such as anorexia, weight loss, fatigue or hypertrophic pulmonary osteo-arthropathy; no anatomic evidence, by X-ray or biopsy, of tumor spread outside the lungs; and pulmonary comorbidity. For further details, see text.

Although the numbers are small and would have to be much larger for convincing statistics, these small numbers of data have much greater clinical pertinence for a patient in this resemblance group than vast numbers, derived from general statistics about lung cancer, that contain none of the cited clinical specifications.

J. Individual Descriptions

After the survival tabulations are completed, the clinician may want to know more details about the library patients who form a resemblance group, so that he can make more exact comparisons between his current patient and those in the resemblance group. To provide these details, the computer can translate the "machine-readable" data into ordinary English to be read by the clinician. A sample of a primitive format appears in Figure 6 (see next page). If more detail is desired, the original man-made chart can be found from its reference in the code number listed for the patient.

V. SUMMARY

In this research project, a computer program has been developed for a clinician's use in searching the stored background of "clinical experience" to make quantitative estimations of prognosis for a new patient.

Before the project began, data for the complete clinical course of 678 patients with primary lung cancer had been obtained and coded in computer-readable categories. These data, which act as the stored "library" of "clinical experience", include the following information for each patient: demographic, clinical, and paraclinical descriptions of the patient's condition before therapy; diverse decisions of management; the administered treatment; and post-therapeutic events for at least five years after treatment.

PATIENT NO. 406

THE PATIENT'S CODE NUMBER WAS 504.

THE PATIENT'S AGE WAS 49.

THE SEX WAS MALE.

SMOKING HABITS INCLUDED CIGARETTES.

CLINICALLY THE PATIENT DEMONSTRATED RECENT COUGH,RECENT DYSPNEA,CHILLS,FEVER,OR SWEATS,SINGLE PULMONARY INFECTION, ANOREXIA,WEAKNESS,FATIGUE,OR MALAISE,AND PAIN OR PATHOLOGICAL FRACTURE SYMPTOMS IN THORACIC BONE.

XRAYS GAVE EVIDENCE OF PANCOAST TUMOR,AND PRIMARY IN THE RIGHT LUNG.
THE WORKUP INCLUDED POSITIVE PARA-THORACIC NODE BIOPSY.
XRAY FURTHER DESCRIBED THE LOCATION OF THE PRIMARY AS MIDDLE LOBE.

THE HISTOLOGIC CODE NO. FOR THE TUMOR WAS 15.

MISCELLANEOUS FINDINGS INCLUDED CHRONIC DYSPNEA,AND ACTIVE TBC.

SURGERY WAS CONTRAINDICATED BECAUSE DISSEMINATION OF THE TUMOR.

PRETREATMENT INTERVAL (IN MONTHS) EQUALLED 24.0.

THE CODE FOR THE FIRST THERAPEUTIC DECISION = 5.

THE CODE FOR THE PAIRED 1ST THERAPY OR PAIRED REFUSAL EQUALLED 0.

SURVIVAL TIME IN MONTHS EQUALLED 7.6.

Figure 6

A computer-generated summary of the patient's data. For further details, see text.

The newly developed computer procedure has been implemented and tested using an interactive, time-shared, "conversational mode" computer system. At the typewriter "terminal", the physician "reads in" the characteristics of the new patient by either answering a prepared, branched set of questions or by entering the data in a pre-coded format. The computer then determines how many previous patients in the "library" are identical to this new patients in each of 14 main "summary properties", and in the "resemblance group" who had all of the same values in their "chart" of those properties.

In this first "chart", the number of completely identical patients in the resemblance group is seldom large enough to warrant prognostic decisions. By using a series of operational commands that are easily entered at the "terminal", the physician can then reduce the specificity of "resemblance". He can drop or add certain properties, convert some into various "ranges" and "unions", or "merge" or "spread" properties upward or downward in their pre-arranged locations on a "hierarchical tree". After each such alteration of the "chart" of the resemblance group, the computer can be requested to indicate the number of previous patients who had the specified similarities.

When the resemblance group seems large enough, the computer can be asked to print out the different forms of treatment given to those patients, and the subsequent outcome. As new patients are followed, their results can augment the stored information in the "library".

The new computer procedure does not create a prognostic prediction, since it requires that the physician use his own judgment in making decisions about the specificity of resemblance, the size of the numbers that seem meaningful, and the interpretation of the results. The main role of the procedure, however, is to allow previous "clinical experience" to be made

available and displayed, promptly and effectively, with documented, quantified details. It enables the decisions of therapeutic management, which formerly depended on uncertain recollections and anecdotal intuitions, to receive the enumerated precision necessary for clinical science.

REFERENCES

1. Feinstein, A.R., Pritchett, J.A. and Schimpff, C.R. The Epidemiology of Cancer Therapy. III. The management of imperfect data. Arch. Intern. Med. 123: 448-461 (April), 1969.

2. Feinstein, A.R., Pritchett, J.A. and Schimpff, C.R. The Epidemiology of Cancer Therapy. IV. The extraction of data from medical records. Arch. Intern. Med. 123: 571-590 (May), 1969.

3. Feinstein, A.R. and Koss, N. Prognosis by computer: Based on storage and retrieval of data for 5-year clinical course of 691 cases of lung cancer. J. Clin. Invest. 45: 1007 (June), 1966. [Abstract; Complete reports submitted for publication.]

4. Scherr, A.L. An Analysis of Time-Shared Computer Systems. MIT Press, Cambridge, Mass., 1967.

5. CYTOS Users Manual. 2nd Ed. Yale Computer Center Mimeograph, Yale University, July, 1968.

6. IBM Manual. 6th Ed. IBM System-360. FORTRAN IV Language. IBM Corp. Programming Systems Publication, 1271 Avenue of the Americas, New York. 1965-1966, Form C28-6515-5, File S360-25.

7. Feinstein, A.R., Pritchett, J.A. and Schimpff, C.R. The Epidemiology of Cancer Therapy. II. The clinical course: data, decisions, and temporal demarcations. Arch. Intern. Med. 123: 323-344 (Mar.), 1969.