Best Cost-Benefit Clinical Data Next to Investigate at Each Diagnostic Step

CARLOS FEDER, MD, Internal Medicine; TOMAS FEDER, PhD, Computer Sciences

In 2008, we devised a computer program capable of diagnosing diseases in actual patients, as opposed to previous existing programs only usable as training or educational tools for internal medicine. Essentially every computer diagnostic program, provided with patient's clinical data (symptoms, physical signs, and results of tests or diagnostic procedures), retrieve quite a long list of potential diagnoses that integrate a differential diagnosis list. To determine which of these diagnoses correspond to the disease or diseases afflicting an actual patient, it is necessary to calculate for each of them the probability of achieving this goal, confirming one or more and ruling out the remaining. An accurate and efficient instrument to calculate the mentioned probabilities is our novel mini-max procedure, superior to Bayes formula and other methods used to this purpose. The mini-max procedure (core of our diagnostic system) and its several properties, among them discriminating competing and concurrent diagnoses, have been explained in detail in our previous publications [1], [2]. An extremely important function of our program is to recommend at each step of the diagnostic quest the best cost-benefit clinical data next to investigate in the patient, to reach a successful end of the diagnostic quest with the smallest overall cost and greatest efficiency. Here again, our mini-max procedure is most useful, coupled with other novel devises, to minimize prescription of futile tests and procedures, in favor of those that are more accurate and efficient, saving valuable medical resources, while benefiting patients, physicians, nurses, health insurance companies, and the entire medical establishment. Such program is also expected to reduce unjustified malpractice suits.

The best cost-benefit clinical datum next to investigate for presence or absence in a patient is an important function that can substantially shorten and reduce the cost of a diagnostic quest by precluding investigation of futile clinical data. This has important socioeconomic implications, especially in this era of managed care, when insurance companies curtail tests and procedures, and when physicians are rated by their proficiency in ordering tests in general.

Computers are faster and more accurate than the human brain in selecting the most convenient clinical datum next to investigate for presence or absence in the patient at each diagnostic step, based on probabilistic calculations.

Because the term best cost-benefit clinical datum next to investigate in a patient is lengthy, we shorten it to *best cost-benefit clinical datum*.

Initial clinical data collection was achieved during the medical history and physical examination. We accepted whatever clinical data were revealed, without considering their rule-in or rule-out power. Subsequent clinical data collection is more selective, because we have a differential diagnosis list and a better-structured diagnostic process that enables to apply statistical and probabilistic concepts, and choose the best cost-benefit clinical datum, based on cost, positive predictive value (PP value), and sensitivity (S).

A practical way to calculate S of a specific clinical datum for a given disease is to determine statistically the fraction of patients afflicted by this disease who manifest the clinical datum:

$$Sensitivity = \frac{Number\ of\ disease\ cases\ manifesting\ the\ clinical\ datum}{Total\ number\ of\ disease\ cases}$$

In our program we calculate PP value with the following equation:

$$PP\ value\ i = \frac{Si}{S1 + \cdots + Si + \cdots + Sn} \quad (1)$$

Where PP value i = positive predictive value of the clinical datum for the disease i under consideration

Si = sensitivity of the clinical datum for the disease i under consideration

S1 ... Sn = sensitivities of the same clinical datum for corresponding diseases

In our context *cost* to obtain a clinical datum involves not only expense, but also risk and discomfort resulting from the required test or procedure. We assign to each clinical datum one of four overall cost categories: no cost (clinical data typically obtained through medical history and physical examination), small cost (*e.g.*, obtained through routine laboratory analysis, ECG, and other ancillary studies), intermediate cost (*e.g.*, colonoscopy, lymph node excision biopsy), and great cost (*e.g.*, liver biopsy, laparoscopy, laparotomy). *Benefit* of a clinical datum is measured by the magnitude of change it produces in the probability (P) of the respective diagnosis, in turn depending on the magnitude of PP value of clinical data present, which increase P, and the magnitude of S of clinical data absent, which decrease P. The mini-max procedure calculates these P for corresponding diagnoses. Detailed explanations and examples can be found in our previous publication [2].

A recommended best cost-benefit clinical datum can be evaluated—before actually accomplishing the corresponding test or procedure—by *virtually* considering it either present or absent, while observing how much it improves the diagnostic outcome.

The best cost-benefit clinical datum function enables us to predict which new clinical datum will most increase or decrease the total probability (P) of a diagnosis, reducing the number of clinical data required to achieve a final diagnosis.

To select a best cost-benefit clinical datum, several steps must be followed:

Step 1. Select clinical data not yet investigated in the patient

The program examines *every* diagnosis in the differential diagnosis list and selects from its respective disease model in the database (disease listing all potential clinical data it can manifest) all clinical data *not yet investigated*. These clinical data differ from those initially collected; they are expected to be numerous because each disease model will contribute many new clinical data. However, only clinical data of appropriate cost and either of great PP value or great S need be investigated for presence or absence.

Step 2. Organize clinical data not yet investigated according to cost category, diagnosis, PP value, and S

Clinical data not yet investigated are organized according to three hierarchical levels (Fig. 1).

The *first level* is **COST CATEGORIES**, comprising four categories: none, small, intermediate, and great. The *second level* is **DIAGNOSES**, comprising all diagnoses in the differential diagnosis list,

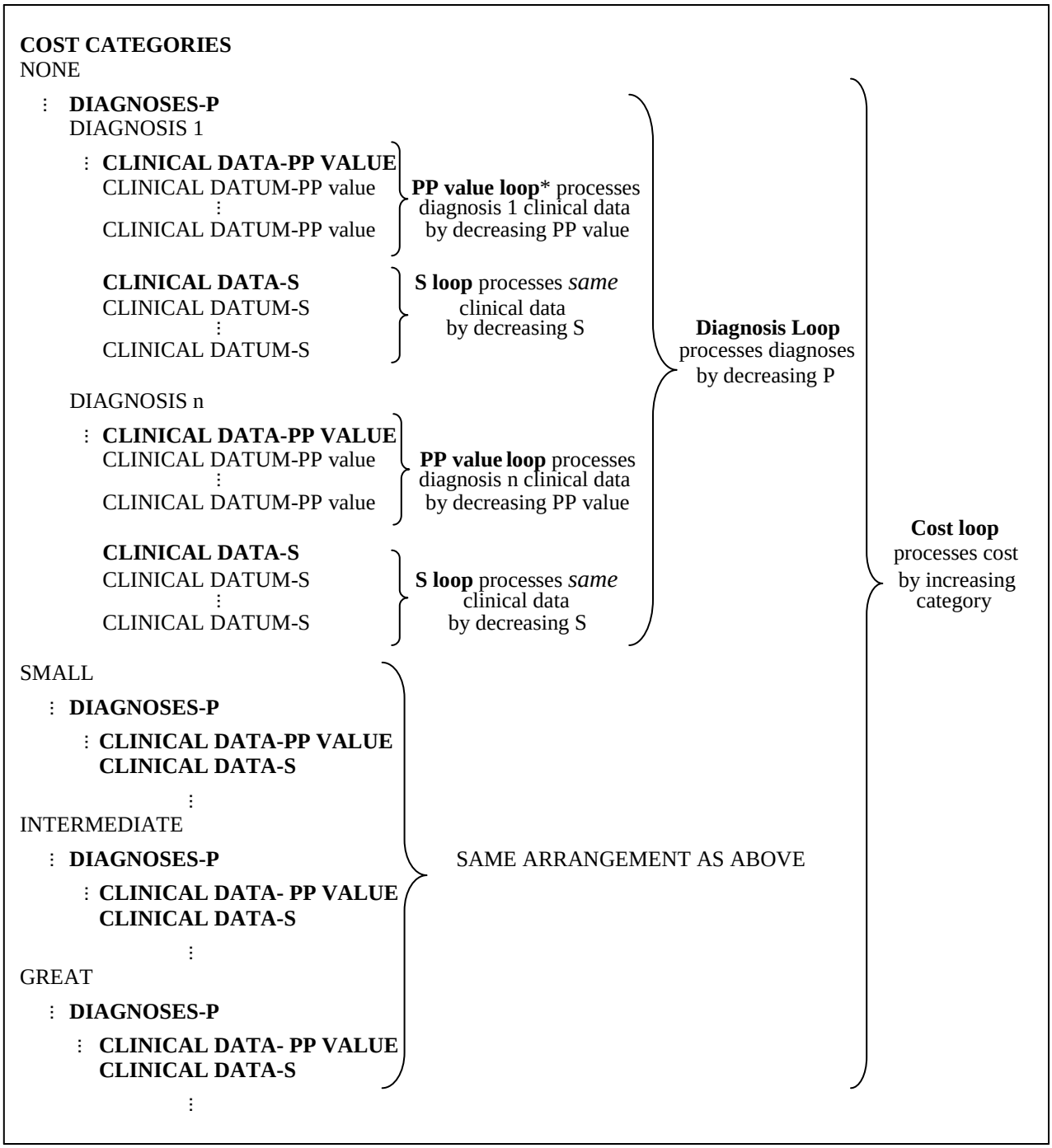

FIGURE 1. Nested loops for selecting best cost-benefit clinical datum.
* A loop is a program fragment that is iterated with fresh operands until a terminal condition is attained; it will be discussed later.

identically repeated in each cost category, in order of decreasing P value. The *third level* is **CLINICAL DATA**, comprising two lists that we call **PP VALUE LIST** and **S LIST** containing only those clinical data that have a cost similar to the corresponding cost category. Both lists contain the same clinical data, but sorted according to decreasing PP value and decreasing S value respectively, and consequently these clinical data are shown with different sequence in each list.

Step 3. Recommend a new clinical datum as best clinical datum assuming it present

Our program creates for each clinical datum present entered in the computer what we call a *clinical datum list* that has for heading this clinical datum and lists all diagnoses capable of manifesting such clinical datum.

Our diagnostic algorithm considers that the greatest PP value of related clinical data present supporting a specific diagnosis better represents the P of this diagnosis than any arithmetical combination of the individual PP values.

For each diagnosis in the differential diagnosis list, our program looks in the entire set of clinical datum lists and selects the greatest PP value that supports *this* diagnosis. The selected greatest PP value equals the P of this diagnosis.

$$P = \max\ (PP\ value\ 1\ \ldots\ PP\ value\ n) \qquad (2)$$

Where P = probability of the diagnosis under consideration

max = maximum of

PP value $_1$… PP value $_n$ = positive predictive values of clinical data present, that support the diagnosis under consideration

The algorithm then iterates the same routine to determine the P of each diagnosis in the differential diagnosis list.

Example: a patient presents with cough, hemoptysis, dyspnea, expectoration, and *Mycobacterium tuberculosis* (*Mycobacterium TB*) in sputum. Five clinical datum lists are generated:

Cough	**S**	**PP value**
Pulmonary tuberculosis	0.80	0.276
Pulmonary embolism	0.50	0.172
Bronchiectasis	0.90	0.310
Lung cancer	0.70	0.241

Hemoptysis	**S**	**PP value**
Pulmonary tuberculosis	0.40	0.222
Pulmonary embolism	0.60	0.333
Bronchiectasis	0.30	0.167
Lung cancer	0.50	0.278

Dyspnea	**S**	**PP value**
Pulmonary tuberculosis	0.20	0.148
Pulmonary embolism	0.50	0.370
Bronchiectasis	0.05	0.037
Lung cancer	0.60	0.444

Expectoration	**S**	**PP value**
Pulmonary tuberculosis	0.80	0.417
Pulmonary embolism	0.02	0.010
Bronchiectasis	0.90	0.469
Lung cancer	0.20	0.104

***Mycobacterium* TB**	**S**	**PP value**
Pulmonary tuberculosis	0.70	1.000
Pulmonary embolism	0.00	0.000
Bronchiectasis	0.00	0.000
Lung cancer	0.00	0.000

S values in the above example are for demonstration purposes only and do not represent actual statistics. PP values were calculated by applying equation 1 to these S values. We assume that only the four listed diagnoses exist and that any of them could account for the five clinical data. Highlighted values refer to clinical data that are not elements of a specific disease model; accordingly, their S and PP value values equal 0. Such clinical data have no influence on calculated probabilities.

In the entire set of clinical datum lists, the greatest PP value for pulmonary tuberculosis is 1.000 and equals P for this diagnosis. Similarly, 0.370 for pulmonary embolism, 0.469 for bronchiectasis, and 0.444 for lung cancer

If clinical data absent were not considered so far, then these greatest values would represent P of the diagnoses. But when clinical data absent were previously considered, which may reduce these P, the mini-max procedure must be applied to calculate resulting P (for details of how mini-max procedure processes information and calculates P of diagnoses, please refer to our previous publication [2]; here we will summarize only the basics).

Mini-max procedure requires to create clinical data pairs, each comprising one clinical datum present and one absent; applying equation 3 (see below) to each of these pairs yields a *partial probability* (partial P) for the corresponding diagnosis. All the partial P are transferred to the so called mini-max tables—one table for each diagnosis—(see example, Table 1, below), and the partial P that is at the same time the smallest in its row and the greatest in its column is called the *determining partial P* because it equals the total P of the diagnosis. The specific clinical data pair that yielded the determining partial P is called the *determining clinical data pair*.

$$Pi = \frac{PP\ value\ i\ (1 - Si)}{PP\ value\ 1\ (1 - S1) + \ \ldots\ + \ PP\ value\ i\ (1 - Si) + \ \ldots\ + \ PP\ value\ n\ (1 - Sn)} \qquad (3)$$

Where P_i = probability of a diagnosis i

$PP\ value_i$ = positive predictive value of the clinical datum present

S_i = sensitivity of the clinical datum absent

$PP\ value_1 \ldots PP\ value_i \ \ldots\ PP\ value_n$ = positive predictive value of the same clinical datum present for each respective diagnosis in the differential diagnosis list

$S_1 \ldots S_i \ldots S_n$ = sensitivity of the clinical datum absent for each respective diagnosis in the differential diagnosis list

Mini-max table for lung cancer

LUNG CANCER	PP value = partial P before considering clinical data absent	Partial P with Cavity absent S = 0.3	Partial P with Fever absent S = 0.1	MINIMUM VALUE IN EACH ROW
Cough present	0.241	0.231	0.297	0.231
Hemoptysis present	0.278	0.254	0.349	0.254
Dyspnea present	0.444	*0.402*	0.540	0.402
Expectoration present	0.104	0.109	0.135	0.104
MTb present	0.000	0.000	0.000	0.000
MAXIMUM VALUE IN EACH COLUMN	0.444	0.402	0.540	**Total P = 0.402**

TABLE 1. The total probability of **lung cancer** at this diagnostic step is the maximum value (**0.402**) in the last column.

To calculate P of diagnoses, when clinical data absent are involved, the algorithm moves to the lowest as-yet-unprocessed COST category, selects the as-yet-unprocessed DIAGNOSIS with greatest P, and

from the corresponding PP VALUE LIST, selects the as-yet-unprocessed clinical datum with the greatest PP value. This PP value then is compared to the PP value of the clinical datum present in the *current* determining clinical data pair. An example of only one clinical data pair, from all possible combinations, and results of equation 3 applications follows:

Dyspnea-Cavity	**PP value**		**S**		**numerators**		**denominator**		**Partial P**
Pulmonary tuberculosis	0.148	×	(1-0.60)	=	0.059	÷	0.774	=	0.077
Pulmonary embolism	0.370	×	(1-0.00)	=	0.370	÷	0.774	=	0.478
Brochiectasis	0.037	×	(1-0.10)	=	0.033	÷	0.774	=	0.043
Lung cancer	**0.444**	×	(1-**0.30**)	=	0.311	÷	0.774	=	**0.402**
					0.774				1.000

The greatest PP value of current clinical data present (0.444) appears in the bottom cell of the second column of the mini-max table for this diagnosis (*e.g.*, lung cancer), and equals the current P of the diagnosis *before* processing clinical data absent. New clinical data with equal or smaller PP value can be disregarded because—even if present—they will not change the current P of this diagnosis; the algorithm moves to Step 4. When the new clinical datum has a PP value that exceeds the current P of the diagnosis before considering clinical data absent (bottom cell of second column), the algorithm recommends it as best cost-benefit clinical datum. The user then verifies whether this clinical datum is absent or present. When this clinical datum is absent, it is disregarded, because if able to change the total P of the diagnosis, it will be detected by the S loop at the next Step 4, which processes clinical data assumed absent. When the recommended best cost-benefit clinical datum is present (*e.g.*, pulmonary mass evidenced by X-ray plain films) it is entered in the computer, the program is iterated, applying the entire mini-max procedure from start. A new clinical datum list headed by this datum and listing all diagnoses able to manifest it is generated, and the partial P of the diagnosis *before* considering clinical data absent assumes the PP value of this new datum. To recalculate the total P of the diagnosis *after* considering clinical data absent, several new clinical data pairs, combination of the new best cost-benefit clinical datum present with each previous clinical datum absent, are generated and the partial P values for the diagnoses are calculated with equation 3. A new row with these values is inserted in each mini-max table and the total P values of the corresponding diagnoses are established.

Mini-max table for lung cancer when pulmonary mass is present

LUNG CANCER	PP value = P before considering absent clinical data	Cavity absent S = 0.3	Fever absent S = 0.1	MINIMUM VALUE IN EACH ROW
Cough present	0.241	0.231	0.297	0.231
Hemoptysis present	0.278	0.254	0.349	0.254
Dyspnea present	0.444	0.402	0.540	0.402
Expectoration present	0.104	0.109	0.135	0.104
MTb present	0.000	0.000	0.000	0.000
Pulmonary mass present	0.857	*0.875*	0.925	0.857
MAXIMUM VALUE IN EACH COLUMN	0.857	0.875	0.925	**Total P = 0.857**

TABLE 2. MTb, *Mycobacterium tuberculosis*

Example (continues):

Assume we need to know whether the clinical datum pulmonary mass as evidenced by X-ray plain films, when *present*, can *increase* the current total P of lung cancer (see mini-max table for lung cancer above). The PP value of a pulmonary mass for lung cancer, stored in the disease model in the database, is 0.857. Mini-max procedure states that the greatest PP value of the clinical data supporting a diagnosis equals the P value of this diagnosis:

$P_{\text{lung cancer}} = \max(PP\ value_{\text{cough}}, PP\ value_{\text{hemoptysis}}, PP\ value_{\text{dyspnea}}, PP\ value_{\text{expectoration}}, PP\ value_{\text{MTb}}, PP\ value_{\text{mass}})$

$= \max(0.241, 0.278, 0.444, 0.104, 0.000, 0.857) = 0.857$

Accordingly, the previous P of lung cancer (0.444), *before considering clinical data absent* (bottom cell of second column in the lung cancer mini-max table), is increased to its new value of 0.857. The algorithm then creates a new clinical datum list, headed by the clinical datum and listing the diagnoses capable of manifesting this datum:

Pulmonary mass	**S**	**PP value**
Lung cancer	0.9	0.857
Pulmonary tuberculosis	0.1	0.095
Pulmonary embolism	0.05	0.048
Bronchiectasis	0.0	0.000

To determine the total P of the diagnosis, *after considering clinical data absent*, the algorithm creates two new clinical data pair tables and calculates the partial P values for lung cancer, applying equation 2:

Mass-Cavity	**PP value**		**S**				**Partial P**
Lung cancer	0.857	×	(1-0.30)	=	0.600	÷ 0.686	= **0.875**
Pulmonary tuberculosis	0.095	×	(1-0.60)	=	0.038		
Pulmonary embolism	0.048	×	(1-0.00)	=	0.048		
Bronchiectasis	0.000	×	(1-0.10)	=	0.000		
					0.686		

Mass-Fever	**PP value**		**S**				**Partial P**
Lung cancer	0.857	×	(1-0.10)	=	0.771	÷ 0.833	= **0.925**
Pulmonary tuberculosis	0.095	×	(1-0.70)	=	0.028		
Pulmonary embolism	0.048	×	(1-0.30)	=	0.034		
Bronchiectasis	0.000	×	(1-0.00)	=	0.000		
					0.833		

In the mini-max table for lung cancer, the algorithm creates a new row that shows these partial P values; then, the total P of this diagnosis *after* considering clinical data absent is calculated:

The total P of **lung cancer** at this diagnostic step is the maximum value (**0.857**) in the last column. Because of pulmonary mass *present* in chest X-ray plain films, the total P of lung cancer increased from 0.402 (before the new row *Pulmonary mass present* was introduced) to 0.857. In this particular case, the clinical data absent did not reduce P of lung cancer.

Each new best cost-benefit clinical datum present creates a new clinical datum list (pulmonary mass in our previous example) that includes the diagnosis from which disease model it was selected (lung cancer). This diagnosis appears in some or all previous clinical datum lists because it originated the search for the new clinical datum; the latter just increases the number of clinical data that support this diagnosis and also its P. Some of the new clinical datum lists created may include previously unlisted diagnoses (*e.g.*, hydatid cyst; see Fig. 2) that also may manifest this clinical datum (pulmonary mass). When this occurs, such new diagnoses will not have clinical data in common with any previous diagnosis because they were not included in previous clinical datum lists; accordingly, previous and new diagnoses, if confirmed as final, must be concurrent. New clinical datum lists select new diagnoses; these, in turn, select new clinical data. At first thought, this cycle may seem to iterate indefinitely until the universe of clinical data is exhausted. In reality, this does not occur, because a single patient cannot manifest all clinical data. At some point, the newly recommended best cost-benefit clinical datum will simply be absent and will not

create a new clinical datum list, aborting the cycle; still, it must be investigated so as to confirm its absence. Neither can a patient be afflicted by a multitude of concurrent diseases. Another factor limiting the number of diagnoses is that a best cost-benefit clinical datum is selected for its great PP value and accordingly is either pathognomonic for a single diagnosis or supportive of only a few diagnoses.

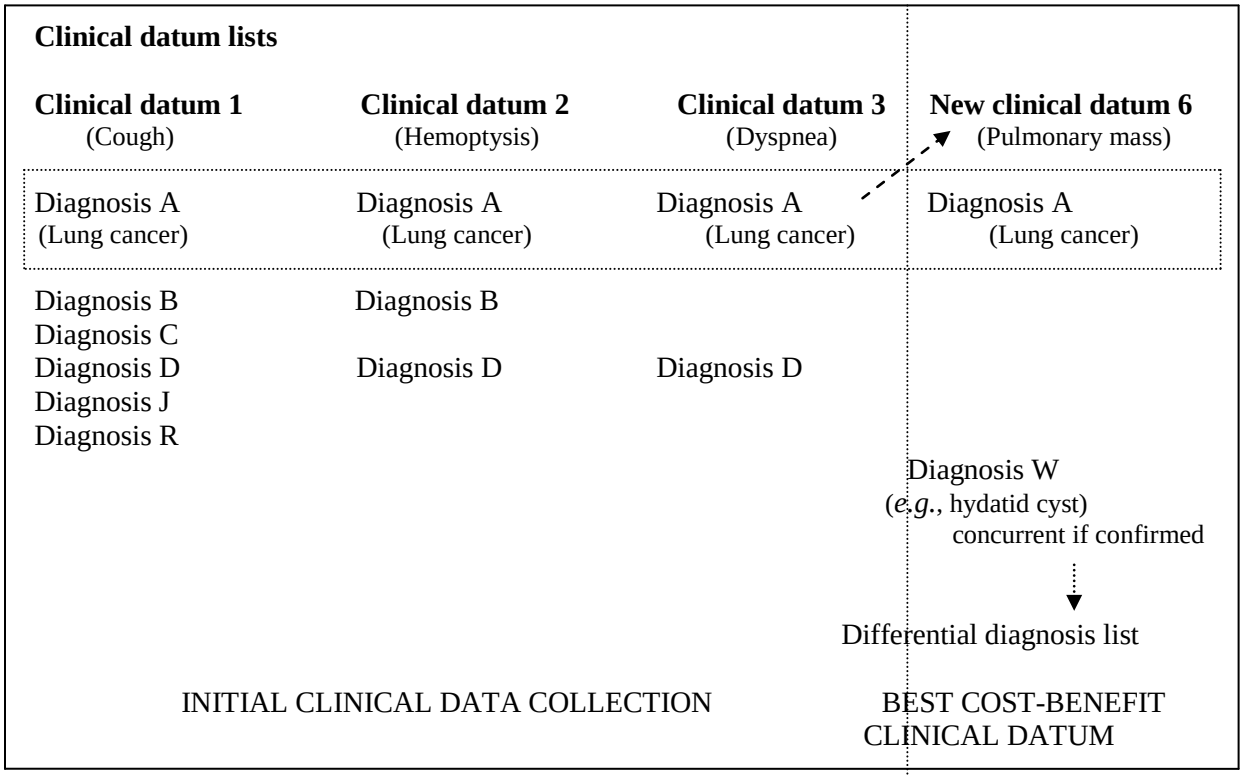

FIGURE 2. To increase probability of diagnosis A, supported by clinical data 1, 2, and 3 (dotted box), clinical datum 6 is recommended as best cost-benefit clinical datum. Clinical datum 6, confirmed present in the patient, creates clinical datum list 6 that includes diagnosis A from which PP list clinical datum 6 was selected; this increases P of diagnosis A and also the number of supporting clinical data. Previously unlisted diagnosis W, also supported by clinical datum 6, is included in the differential diagnosis list as a new diagnosis; if it reaches confirmation threshold, it will become a *concurrent* final diagnosis.

Step 4. Recommend a new clinical datum as best clinical datum assuming it absent

Typically, the greater the S of a clinical datum absent, the more it decreases the total P of a diagnosis. However, with the mini-max procedure, a clinical datum absent can occasionally increase total P. It also can occur that a clinical datum with a smaller S can decrease total P more than *another* clinical datum absent with a greater S. This paradox is due to broken monotony* or a particular interaction of partial P values in the corresponding mini-max table. When broken monotony occurs, the resultant P is only slightly different from the expected P and does not compromise the efficiency of our program.

*** Broken monotony**

Typically, the partial P values in the rows of the mini-max table present a monotone relation; this means that when in one row the partial P value increases or decreases from one cell to the next, in the other rows the changes occur in the same direction. However, sometimes this monotone relation is broken. This is due to the especial interrelation among the diverse S and PP values in the clinical data pair tables, and has several consequences; however, our new diagnostic program ignores broken

monotony, and proved to remain accurate and efficient. The interested reader will find more details on this subject in our previous publications [1], [2].

To verify whether a new clinical datum absent will indeed decrease the total P, the mini-max procedure needs not always be applied in its entirety; to save computer time (although this is unlikely to be critical, considering the great speed of today's computers), we devised the 3-Step method. Our current diagnostic program does not apply the 3-Step method because it proved to be unnecessary, at least for the limited number of diseases we are currently processing; however, with a database including all known diseases, this method might be significantly computer time saving. The 3-Step method is described with examples in our previous book [1] to which we refer the interested reader.

To process clinical data absent, the algorithm moves now to the S LIST, in the same COST category and DIAGNOSIS in which a new best cost-benefit clinical datum assumed present was processed involving PP value list. From this S LIST, it selects the as-yet-unprocessed clinical datum with the greatest S value. This S value then is compared to the S value of the clinical datum present in the *current* determining clinical data pair. New clinical data with equal or smaller S value can be disregarded because—even if absent—they are assumed not to change the current P of this diagnosis. When the new clinical datum has an S value that exceeds the current S in the *current* determining clinical data pair, it replaces the current S. Equation 3 is applied to this new clinical data pair and if resulting P of corresponding diagnosis is smaller than the current one, the algorithm recommends the clinical datum as best cost-benefit clinical datum. This processing represents a simplified mini-max procedure, equivalent to calculate only *one* cell [2] (expected new determining partial P) in the new inserted absent data column of the mini-max table, ignoring potential broken monotony. The user then verifies whether this clinical datum is absent or present.

When this clinical datum is present, it is disregarded, because if able to change the total P of the diagnosis, it would have been detected by the PP value loop at previous Step 3, which processes clinical data assumed present. When the recommended best cost-benefit clinical datum is absent it is entered in the computer, the program is iterated and equation 3 is now applied to all clinical data pairs possible to create, calculating the new determining partial P with the complete mini-max procedure that will fill in *all* the cells of the mentioned column, yielding a more accurate P of the diagnosis. If this new determining partial P is smaller than the current P it will replace this current value; otherwise, it will be disregarded.

Disregarding best cost-benefit clinical data selected from the S list, because they are present, same as disregarding data selected from the PP value list, because they are absent, can be questioned. These data, although not expected to change diagnosis P, may however produce a slight improvement of this P, now calculated with all cells of the mini-max table as opposed to only one cell in the new column. At least, the mentioned disregarded clinical data should be entered in the computer as a reminder that they were already processed and will not be recommended again.

Example: Assume we need to know whether the clinical datum pulmonary mass as evidenced by chest X-ray plain films, if *absent*, can *decrease* the total P of lung cancer, that was 0.402 before considering this datum present in previous example (see lung cancer table 1, above). The P of lung cancer *before* considering clinical data absent was 0.444 (equals greatest PP value in the second column, corresponding to dyspnea present); S of pulmonary mass on X-ray films for lung cancer is 0.9, as shown in the corresponding S LIST.

To calculate resultant P that a best cost-benefit clinical datum assumed absent confers to each respective diagnosis, we calculate new partial P only for one cell in the new column of corresponding mini-max table; but as we must achieve this for all diagnoses in the differential diagnosis list, we apply equation 3 to each of these diagnoses in the clinical data pair Dyspnea-Mass:

Dyspnea-Mass	**PP value**		**S**				**Partial P**
Pulmonary tuberculosis	0.148	×	(1- 0.10) = 0.133	÷	0.566	=	0.235
Pulmonary embolism	0.370	×	(1- 0.05) = 0.352	÷	0.566	=	0.622
Bronchiectasis	0.037	×	(1- 0.00) = 0.037	÷	0.566	=	0.065
Lung cancer	0.444	×	(1- 0.90) = 0.044	÷	0.566	=	0.078
			0.566				1.000

For this example we show equation 3 below, applied only to the diagnosis lung cancer and the corresponding mini-max table:

$$P_{\text{lung cancer}} = \frac{\text{PP value}_{\text{lung cancer}} (1- S_{\text{lung cancer}})}{\text{PPvalue}_{\text{lung cancer}} (1-S_{\text{lung cancer}}) + \text{PPvalue}_{TB} (1-S_{TB}) + \text{PPvalue}_{\text{embolism}} (1-S_{\text{embolism}}) + \text{PPvalue}_{\text{bronchiectasis}} (1-S_{\text{bronchiectasis}})}$$

$$P_{\text{lung cancer}} = \frac{0.444 (1- 0.90)}{0.444 (1- 0.90) + 0.148 (1- 0.10) + 0.370 (1- 0.05) + 0.037 (1- 0.00)}$$

$$= \frac{0.044}{0.044 + 0.133 + 0.352 + 0.037} = 0.078$$

Mini-Max Table for Lung Cancer when pulmonary mass is absent

LUNG CANCER		PP value = partial P before considering absent clinical data	Cavity absent S = 0.3	Fever absent S = 0.1	Mass absent S = 0.9	MINIMUM VALUE IN EACH ROW
Cough	present	0.241	0.231	0.297		
Hemoptysis	present	0.278	0.254	0.349		
Dyspnea	present	0.444	0.402	0.540	0.078	0.078
Expectoration	present	0.104	0.109	0.135		
MTb	present	0.000	0.000	0.000		
MAXIMUM VALUE IN EACH COLUMN		0.444	0.402	0.540		**Total P = 0.078**

TABLE 3. MTb, *Mycobacterium tuberculosis*

According to the above explanation, as expected, total P (0.078) is smaller than the current P (0.402), "Mass absent" is recommended as best cost-benefit clinical datum. If this datum is confirmed absent, entered in the computer, and processed, the complete mini-max procedure is applied and all cells of the mini-max tables, including the new column, will be filled with partial P. This requires the creation of all corresponding clinical data pairs; we show only the results in the Table 4.

Mini-Max Table for Lung Cancer

LUNG CANCER	PP value = partial P before considering absent clinical data	Cavity absent S = 0.3	Fever absent S = 0.1	Mass absent S = 0.9	MINIMUM VALUE IN EACH ROW
Cough present	0.241	0.231	0.297	0.032	0.032
Hemoptysis present	0.278	0.254	0.349	0.039	0.039
Dyspnea present	0.444	0.402	0.540	*0.078*	0.078
Expectoration present	0.104	0.109	0.135	0.012	0.012
MTb present	0.000	0.000	0.000	0.000	0.000
MAXIMUM VALUE IN EACH COLUMN	0.444	0.402	0.540	0.078	**Total P = 0.078**

TABLE 4. MTb, *Mycobacterium tuberculosis*

Mini-Max Table after absence of Mass was confirmed

In summary, to select the best cost-benefit clinical datum to investigate next, the algorithm loops at three nested levels (Fig. 1, above): outer, intermediate, and inner. (1) The outer *cost loop* processes clinical data not yet investigated in order of increasing cost category: none, small, intermediate, and great. (2) Within each cost category, the intermediate *diagnosis loop* processes the diagnoses of the differential diagnosis list in order of decreasing P because those with greatest P values, are the best candidates for a final diagnosis, and can sooner conclude the diagnostic quest. (3) The inner *clinical data loop* comprises two sub-loops: the first begins at the top of the PP value list and terminates when no clinical datum exists able to increase current P of the corresponding diagnosis. The second sub-loop begins at the top of the S list and terminates when no clinical datum exists able to decrease current P of the corresponding diagnosis.

All diagnoses in each cost category are similarly processed. The user is prompted each time the loop goes to a greater cost category. The remaining differential diagnoses with their P are displayed and the user is asked whether he wants to proceed in the greater cost category or prefers a deferred diagnosis, diagnosis by exclusion, or empirical treatment [1], [2]. The entire looping process terminates when all final diagnoses are obtained and competing diagnoses are ruled out, the cost of investigating recommended clinical data exceeds the benefit, or all the clinical data able to change P of diagnoses are processed. Clinical data that have the greatest PP value or the greatest S typically involve costly pathological investigations, such as biopsy or even autopsy. To request a biopsy or even an autopsy for a patient with tonsillitis would be crazy. This exaggeration emphasizes the importance of initially considering the cost of a clinical datum, before evaluating its PP value or S. However, in an emergency or when a patient's condition is deteriorating, investigation of confirmatory clinical data of great PP value takes priority over cost.

The recommended best cost-benefit clinical datum could be a common symptom quickly asked or immediately observed by the physician. Should obtaining a clinical datum require an involved test or procedure, the diagnostic process must be interrupted until the result becomes available. The "position of the game board", so to say, must be saved in the computer and opportunely retrieved to continue the "game", because each new clinical datum, with its presence or absence in the patient, sets a new stage for recommending the next best cost-benefit clinical datum. A disease is not a static process. If the clinical

picture changes considerably before obtaining the diagnostic test or procedure result, a new diagnostic process must be accomplished, sometimes from the beginning.

A new best cost-benefit clinical datum, present or absent, changes the P of the diagnosis from which it was selected, and sometimes also some of the other diagnoses that can manifest the same clinical datum. We had some doubts whether these P changes in the other diagnoses were legitimate. However, our program processes all diagnoses simultaneously with the mini-max procedure, calculating from the start P of all diagnoses, each time a new clinical datum, present or absent is entered in the computer. Although a best cost-benefit clinical datum has been selected from a list related to a specific diagnosis, once entered in the computer and processed, the resulting P of every diagnosis in the differential diagnosis list is the correct one, irrelevant from which particular diagnosis in mind it was selected.

COMMENTS

Clinical data—symptoms, physical signs, test results, and diagnostic procedure results—are collected from a patient and matched with diseases models in the database, selecting all diseases that present a match as potential diagnoses. This typically creates a pretty large differential diagnosis list. The next problem is to determine which of these diagnoses are the one or more than one that actually afflict the patient, which requires to calculate the probability of each such diagnoses. Diagnoses that reach a probability of certain empirically determined confirmation threshold are considered final and representing the diseases that afflict the patient, whereas diagnoses that yield a probability below an empirically determined deletion threshold are ruled out. Bayes formula and other mathematical instruments used in previous existing programs do not address satisfactorily this problem; for this reason they are presented as training tools or educational tools rather than efficient diagnostic aids. Most of these programs are not capable to diagnose several diseases afflicting simultaneously a single patient (concurrent diseases), a situation that occurs frequently in complex clinical presentations. A computer program executing our mini-max procedure provided us with a prototype that proved to diagnose accurately and efficiently when challenged with real clinical cases, including concurrent diseases.

Our complete diagnostic program includes several other important functions that are expected to be published in coming papers, and currently described in our book *Computerized Medical Diagnosis: A Novel Solution to an Old Problem* [1] that stresses theoretical and historical issues, and in our recent and more practical book *A Practical Computer Program that Diagnoses Diseases in Actual Patients* [2]. Some of these functions, not discussed in the present paper are:

- Recommendation of a **set** of best cost-benefit clinical data to be investigated simultaneously at each diagnostic step. We realize that such clinical data cannot be recommended sequentially one by one; this would be too time consuming, particularly in emergencies, and would require an excessive number of patient-physician encounters and patient hardship.

- Empirical parameters that enable to reduce the number of recommended clinical data to investigate simultaneously, without compromising the accuracy of the diagnosis, and different output lists of recommended clinical data that facilitate the selection of such set of data based on preferences and diverse medical circumstances.

- Safeguard function that precludes overlooking diagnoses associated with diseases confirmed by our program; this function is based on complex clinical presentation models that include diagnoses related by pathophysiologic mechanisms or statistical correlations.

- Safeguard function that precludes missing a diagnosis (*e.g.*, myocardial infarction), when an important clinical datum (chest pain) is masked by interaction with a concurrent disease (diabetes) or drug (potent analgesic).

The database of our current diagnostic prototype is integrated with 50 diseases models. Our diagnostic system, once implemented with all known diseases and clinical data, is expected to provide invaluable diagnostic benefits to patients, physicians, nurses, health insurance companies, malpractice lawyers, and the entire medical establishment.

CONCLUSIONS

Our algorithm, although somewhat complex, is straightforward, especially when compared to other attempts in this field. It emulates a clinician's diagnostic reasoning. It is logical and mathematically simple. Bayes formula is used with modifications, because it is unable to process properly interdependent clinical data (as are most symptoms) and concurrent diseases. To facilitate implementation and updating of the algorithm, we tend to avoid complicated tools of artificial intelligence, such as causal, hierarchical, and probabilistic trees and networks. The algorithm freely uses heuristic procedures, so as to preclude excessive proliferation of clinical data and diagnoses. It promises to be user friendly because it is expressed in natural language, is rational, and readily understandable. Determination of accurate sensitivity of clinical data and integration of clinical entities into complex clinical presentation models will be labor-intensive. A complete database with all known diseases, clinical data, clinical presentations, and other information can be created; this major task will require a dedicated team of medical specialists.

REFERENCES

[1] FEDER C. Computerized Medical Diagnosis: A Novel Solution to an Old Problem. Infinity Publishing, West Conshohocken, Pennsylvania, 2006

[2] FEDER C, and FEDER T. A Practical Computer Program that Diagnoses Diseases in Actual Patients. Infinity Publishing, West Conshohocken, Pennsylvania, 2008

LEDLEY RS and LUSTED LB. Reasoning Foundations of Medical Diagnosis. Science, 130 (9): 9-21, July 3, 1959

HENRION M, PRADHAN M, DEL FAVERO B, HUANG K, and O'RORKE P: Why is diagnosis using belief networks insensitive to imprecision in probabilities? Twelfth Conference on Uncertainty in Artificial Intelligence, Portland, OR, 446-454. 1996. SMI-96-0637

MYERS JD, POPLE HE, and MILLER RA. INTERNIST: Can Artificial Intelligence Help? In: Connelly, Benson, Burke, Fenderson, eds. Clinical Decisions and Laboratory Use. Minneapolis: University of Minnesota Press, 1982: 251-269

LUSTED LB. Introduction to Medical Decision Making. Springfield, Illinois, Charles C Thomas, 1968

LUSTED LB. Twenty Years of Medical Decision Making Studies. CH1480-3/79/0000-0004$00.75. 1979 IEEE

POPLE HE. Heuristic Methods for Imposing Structure on Ill-structured Problems: The Structuring of Medical Diagnosis. In: Szolovits P, ed. Artificial Intelligence in Medicine, AAAS Symposium Series, Boulder, Colorado: West-view Press, 1982: 119-185

MARTIN J: Computer Data-Base Organization. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1977

CHAKRAVARTY S and SHAHAR Y. A Constraint-Based Specification of Periodic Patterns in Time-Oriented Data. Sixth International Workshop on Temporal Representation and Reasoning (TIME-99), Orlando, FL, 29-40. 1999. SMI-1999-0766

ALTMAN RB. AI in Medicine: The Spectrum of Challenges from Managed Care to Molecular Medicine. AI Magazine 20(3):67-77, 1999. SMI-1999-0770

BLEICH HL. Computer-Based Consultation: Electrolyte and Acid-Base Disorders. The American Journal of Medicine 53: 285-291, November 1972

BLOIS MS, TUTTLE MS, and SHERERTZ DD. RECONSIDER: A Program for Generating Differential Diagnoses. IEEE: 263-268, 1981

CHAKRAVARTY S and SHAHAR Y. Acquisition and Analysis of Periodic Patterns in Time-Oriented Clinical Data. 2000. SMI-2000-0822

DE DOMBAL FT, LEAPER DJ, STANILAND JR, McCANN AP, and HORROCKS JC. Computer-Aided Diagnosis of Acute Abdominal Pain. British Medical Journal 2: 9-13, 1972

ESHELMAN L, EHRET D, McDERMOTT JP, and TAN M (1987.) MOLE: A Tenacious Knowledge Acquisition Tool. International Journal of Man-Machine Studies, 26: 41-54

GENNARI J, MUSEN MA, FERGERSON RW, GROSSO WE, CRUBEZY M, ERIKSSON H, NOY NF, TU SW: The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. 2002. SMI-2002-0943

GINI RA and FEDER C. Informatica Clinica: Presente y Futuro. La Semana Medica (Argentina) 157: 113-128, 1980

GORRY GA, KASSIRER JP, ESSIG A, and SCHWARTZ WB. Decision Analysis as the Basis for Computer-Aided Management of Acute Renal Failure. The American Journal of Medicine 55: 473-484, 1973

GORRY GA, PAUKER SG, and SCHWARTZ WB. The Diagnostic Importance of the Normal Finding. The New England Journal of Medicine 486-489, March 2, 1978

GREENS RA, PELEG M, BOSWALA AA, TU S, PATEL VL, and SHORTLIFFE EH. Sharable Computer-Based Clinical Practice Guidelines: Rationale, Obstacles, Approaches, and Prospects. Medinfo, London, UK, 2001. SMI-2001-0860

HUANG K, HENRION M: Efficient Search-Based Inference for Noisy-OR Belief Networks: Top Epsilon. Proceedings of the Twelfth Conference of Uncertainty in Artificial Intelligence, 325-331. Aug 1996, Portland, OR. SMI-96-0640

ILIAD® 4.5 Diagnostic and Reference Tool for Physicians and Medical Professionals. User Guide. 1998

JAAKKOLA TS, JORDAN MI: Variational Probabilistic Inference and the QMR-DT Network. Sun May 9, 16:22:01 PDT 1999

LUDWIG DW. INFERNET – A Computer-Based System for Modeling Medical Knowledge and Clinical Inference. Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care: 243-249, November 1981

MIDDLETON B, SHWE M, HECKERMAN D, HENRION M, HORVITZ E, LEHMANN H, & COOPER G: Probabilistic Diagnosis using a reformulation of the INTERNIST-1/QMR Knowledge Base II. Evaluation of Diagnostic Performance. Section on Medical Informatics Technical report SMI-90-0329, Stanford University, 1990

MILLER RA, POPLE HE, and MYERS JD. INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. The New England Journal of Medicine, 1982: 468-476

MUSSEN MA, GENNARI JH, and WONG WW: A Rational Reconstruction of INTERNIST-I Using PROTÉGÉ-II. Nineteenth Annual Symposium on Computer Applications in Medial Care. New Orleans, LA, 289-293. 1995. SMI-95-0574

MUSEN MA: Modeling for Decision Support. Section on Medical Informatics. Stanford University School of Medicine. Stanford, CA 94305-5479. SMI-98-0739

NOY NF and MUSEN MA. Ontology Versioning as an Element of an Ontology-Management Framework. Stanford Medical Informatics, Stanford University, 251 Campus Drive, Stanford, CA 94305, USA. SMI-2003-0961. March 31, 2003

PATRICK EA: Decision Analysis in Medicine: Methods and Applications. CRC Press, West Palm Beach, FL, 1979

PERLROTH MG, and WEILAND DJ. Fifty Diseases: Fifty Diagnoses. Year Book Medical Publishers, 1981

POPLE HE, MYERS JD, and MILLER RA: Dialog: A Model of Diagnostic Logic for Internal Medicine. Fourth International Joint Conference on Artificial Intelligence. Tbilisi, Georgia, URRS, 3-8 September 1975, Volume Two. 1975: 848-855

PRADHAN M, DAGUM P: Optimal Monte Carlo Estimation of Belief Network Inference. Proceedings of the Twelfth Conference of Uncertainty in Artificial Intelligence, 446-453. Aug 1996, Portland, OR. SMI-96-0638

SHORTLIFE EH. Computer-Based Medical Consultations: MYCIN. American Elsevier Publishing Company, 1976

SHORTLIFFE EH. The Next Generation Internet and Health Care: A Civics Lesson for the Informatics Community. In C.G. Chute, Ed., 1998 AMIA Annual Symposium, Orlando, FL, 8-14. 1998. SMI-98-0730

SHWE M, MIDDLETON B, HECKERMAN D, HENRION M, HORVITZ E, LEHMANN H, & COOPER G. Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base I. The Probabilistic Model and Inference Algorithms. Methods of Information in Medicine, 30(4):241-255, 1991. SMI-90-0296

SZOLOVITS P and PAUKER SG. Categorical and Probabilistic Reasoning in Medical Diagnosis. Artificial Intelligence. 11: 115-144, 1978

TU SW et al: Modeling Guidelines for Integration into Clinical Workflow. Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. SMI-2003-0970

VAN BEMMEL JH, MUSSEN MA (eds): Medical Informatics. Springer, 1997

WEINSTEIN MC and FINEBERG HV: Clinical Decision Analysis. W. B. Saunders Company, 1980

WEISS S, KULIKOWSKI CA, and SAFIR A. Glaucoma Consultation by Computer. Comput. Biol. Med. 8: 25-40, 1978