The Measurement of Performance in Probabilistic Diagnosis

IV. Utility Considerations in Therapeutics and Prognostics

(From the Department of Public Health and Social Medicine, Erasmus University, Rotterdam, The Netherlands, and the Institute of Human Genetics, University of Copenhagen, Denmark)

J. D. F. Harbema, J. Hilden

It is argued that it is preferable to evaluate probabilistic diagnosis systems in terms of utility (patient benefit) or loss (negative benefit). We have adopted the provision of our overview of scoring performance as if the system were the actual decision-maker (not just an aid to him) and argue that a rational figure of merit is given by the average loss which patients would incur by having the system decide on treatment, the treatment being selected according to the minimum expected loss principle of decision theory.

A similar approach is taken to the problem of evaluating probabilistic systems, but the fundamental differences between treatment selection skill and prognostic skill and their implications for the assessment of such skills are stressed. The necessary elements of decision theory are explained by means of simple examples mainly taken from the acute abdomen, and the proposed evaluation tools are applied to Acute Abdominal Pain data analysed in our previous papers by other (not decision-theoretic) means. The main difficulty of the decision theory approach, viz. that of obtaining good medical utility values upon which the analysis can be based, receives due attention, and the evaluation approach is extended to cover more realistic situations in which utility or loss values vary from patient to patient.

Key-Words: Computer-aided Diagnosis, Probabilistic Diagnosis, Probabilistic Prognostication, Treatment Selection, Utility Assessment, Decision Theory, Prognosis, Scoring Rules, Evaluation of Clinical Skills

DIE LEISTUNGSMESSUNG BEI DER WAHRSCHEINLICHKEITSDIAGNOSE — IV. NÜTZLICHKEITSERwäGUNGEN IN THERAPEUTIK UND PROGNOSTIK

Es wird argumentiert, daß wahrscheinlichkeitstheoretische Diagnosesysteme vorzugsweise nach Nützlichkeit (Vorteil für den Patienten) oder Schaden (Nachteil für den Patienten) ausgewertet werden sollten. Die Autoren benutzen die provisorische Strategie der Leistungsbewertung so, als ob das System der eigentliche Entscheidungsfinder sei (und nicht nur eine Hilfe für diesen), und argumentieren, daß eine vernünftige Maßzahl für den Nutzen durch den durchschnittlichen Schaden gegeben wird, den Patienten erleiden würden, wenn das System über ihre Behandlung entscheiden würde, wobei die Bewertung so ausgewählt wird, daß der Schaden im Sinne der Entscheidungstheorie minimiert wird. Ähnlich wird das Problem der Auswertung wahrscheinlichkeitstheoretischer Prognosen angegangen; aber die grundlegenden Unterschiede zwischen der Geschicklichkeit in der Auswahl der Behandlung und der Prognosestellung und ihre Folgen für die Bewertung solcher Fähigkeiten werden hervorgehoben. Die erforderlichen Elemente der Entscheidungstheorie werden an einfachen Beispielen, vorwiegend aus dem Gebiet des akuten Abdomens, erläutert, und die vorgeschlagenen Auswertungsinstrumente werden auf bereits in früheren Arbeiten mit anderen Methoden nicht entscheidungstheoretischer Natur analysierte Daten über akute Bauchschmerzen angewandt. Dem Hauptproblem bei Anwendung des entscheidungstheoretischen Ansatzes, nämlich dem, gute medizinische Nützlichkeitswerte für die Analyse zu erhalten, wird gebührende Aufmerksamkeit gezollt, und der Auswertungsansatz wird erweitert, um realistischere Situationen zu erfassen, bei denen die Werte für Nutzen und Schaden von Patient zu Patient variieren.

Schlüssel-Wörter: Computerunterstützte Diagnose, wahrscheinlichkeitstheoretische Diagnose, wahrscheinlichkeitstheoretische Prognosestellung, Wahl der Behandlung, Nützlichkeitserwartung, Entscheidungs-Theorie, Prognose, Bewertungsregeln, Bewertung klinischer Fähigkeiten

1. Introduction

During the last two decades much effort has been devoted to constructing diagnostic computer systems which digest symptom data from a patient and reply by assigning a probability to each of several diagnostic alternatives. Despite the ingenuity of the statistical models and algorithms upon which such systems have been built, there has been little general understanding of what constitutes the desirable properties of the output behaviour of such systems. Consequently, the evaluation methods of the past have not been satisfactory. The purpose of the series of papers consisting of Parts I-III & V ([12, 17, 18, 13]) and the present Part IV is to provide a kit of tools which can be used in assessing the performance of such systems.

Part IV deals with methods that try to reflect the consequences of medical action. In this respect it differs radically from its predecessors. There we studied the probabilities by themselves, attention being focused on the ability of the system to assign a large probability to the correct disease alternative. All diseases were treated on an equal footing — at least if equally frequent — and all misdiagnoses carried equal weight. It was only when defining the disease classes that the possibility was left open for the evaluator to lump together conditions that would show a similar reaction to therapy. In this way he might ensure that poor diagnostic advice would only be penalized if its consequences were harmful to the patient in actual clinical practice.

Here, on the other hand, we shall try to place the diagnostic probabilities in their proper clinical value context. In
deciding on action, physicians (should) combine two things: an impression of the patient's condition, i.e., of his or her therapeutic response potential, and also a judgement as to how the patient will value each of the possible outcomes [43]. Thus, a system designed to help physicians form an impression of the patient's condition must be evaluated in terms of consequences of action and their value to the patient.

1. Conceptual Skeleton

Treatment and prognostic action are the two main types of action we have in mind. By prognostic action we mean the act of listing the possible outcomes of the illness and stating how likely or unlikely they appear to be. Amongst other purposes, such a formal prognosis is used to guide the patient. We do not anticipate that machine-made prognoses are handed to the patient, but we do propose that they are evaluated on the basis of their agreement with the actual course of his illness. Similar remarks apply to selection of treatment: The doctor is at liberty to override the therapeutic suggestions of a system. Nevertheless, we assess its ability to make such suggestions by examining what would happen to the patient if he was given the therapy suggested or implied by the system's output. Thus, in either case we evaluate the system as if it were the actual decision-maker and not just an aid. This "as if" approach is justified because a system which is poor as a decision-maker is probably also no good as an aid (see the introduction to Part III for further discussion of this point).

As decision theory forms the basis of the methods to be discussed here, its terminology will be used. Thus, instead of speaking of value or benefit, we shall use the standard terms utility and loss, loss being negative utility, and vice versa. We shall usually speak in terms of loss. It may be helpful for readers who are unfamiliar with decision theory to note from the outset that the loss scale has an arbitrary zero point (origin) and uses an arbitrary unit of measure, e.g., dollars. Any operation entails a risk, especially if it were the actual decision-maker.

However, we shall assume that the unit of measurement is common to all patients in a given medical setting, so that it is meaningful to compare two systems in terms of the average loss across patients that would be incurred by having one system or the other decide. We favour this assumption of comparable loss units, but this is really a matter of opinion. The widely supported dogma that the welfare of every human being is equally precious takes for granted that such a common scale for welfare exists. But it is also true, strictly speaking, that your well-being and mine are fundamentally incommensurable.

1.2. The Plan of the Paper

Sections 2. and 3. contain the necessary preliminaries. In Section 2. we explain some basic concepts of decision theory, and in Section 3. we briefly reintroduce the system evaluation framework and the illustrative data used in the previous parts. (For further details, please see Part I.) Thus equipped, we are able to analyse system performance. Section 4. looks at selection of treatment. Section 5. develops measures of prognostic skill, while in Section 6. we slightly extend our apparatus in order to allow for differences in loss structure between one patient and the next. These two sections can be read independently or skipped.

The Discussion (Section 7.) relates our work to that of others and brings up some intricate points which were beyond the scope of the preceding sections.

2. Loss, Regret, and Choice of Action

We shall consider a well-defined population of clinical decision tasks such as the acute abdomen right on admission. Let \( T_1, \ldots, T_s \) be the \( s \) distinct actions (treatments) under consideration, and \( D_1, \ldots, D_k \) the \( k \) diagnostic alternatives (diseases). For the moment, we will adopt the simplified view that diseases can be defined in such a way that all patients with \( D_j \) will respond to treatment in the same way. This is not to say that they will show exactly the same course of illness, but only that at the point in time when the decision is made the doctor cannot subdivide \( D_j \) cases into groups that require different treatments or have different prognoses. Suppose that the response to treatment \( T_r \) in patients with \( D_j \) has been studied from the viewpoint of suffering, disability, risk of death, family anguish, etc. Then let \( L_{rj} \) be the loss that this response is felt to represent, all aspects considered. Insofar as the response may vary, \( L_r \) will stand for the mean loss. These agreed loss figures form our loss matrix \( L = (L_{rj}) \) of dimensions \( s \times k \).

Example A (Acute Abdominal Pain)

As in the preceding papers, \( D_j \) will stand for non-specific abdominal pain, \( D_j \) for acute appendicitis, and \( D_j \) for all other causes of admission to a surgical ward with acute abdominal pain. The \( D_j \) class is admittedly heterogeneous. We cannot in decency lump together salpingitis and perforated ulcer. Yet we do so here for the purpose of numerical illustration. The reader may, if he wishes, prefer to think of \( D_j \) as acute bowel obstruction. On admission we consider \( s=4 \) treatment regimens:

\[
\begin{align*}
T_1 & = \text{wait and see}, \\
T_2 & = \text{immediate lower abdominal laparotomy}, \\
T_3 & = \text{frequent clinical reassessment (intensi} \\
& \text{fied wait-and-see)}, \\
T_4 & = \text{carry out special diagnostic tests such as } X-
\text{ray before deciding on treatment.}
\end{align*}
\]

The loss matrix may look as in Table 1, part a. In this loss assessment, which should only be regarded as an illustration, a zero loss is incurred when pains subside and the patient is discharged in complete health after a few days. At the other extreme, a (hypothetical) disease-treatment combination that is certain to be fatal would be rated at about 500 loss units, while certain development of non-fatal peritonitis would be rated at about 100 units. Any operation entails a risk, especially in non-\( D_j \) cases. Small penalties have been attached to "unnecessary" investigations and to delayed therapy.

As in Parts I—III, we assume that \( N \) patients named \( H_1, \ldots, H_N \) are diagnosed in the course of the evaluation study. Patient \( H_i \) presents with symptoms \( x_i \), including age, sex, outcome of laboratory tests, etc., which are fed into the diagnostic system. It replies by assigning a probability \( P_i \) to the possibility that \( H_i \) is suffering from \( D_j \). The probability diagnosis matrix \( P = (P_{ij}) \) of dimensions \( N \times k \) comprises the entire set of probability diagnoses produced during the evaluation study. According to the system, the expected loss caused by giving treatment \( T \) to patient \( H_i \) (call it \( E_{it} \)) amounts to

\[
E_{it} = \sum_{j=1}^{k} P_{ij} L_{rj},
\]

i.e., for each \( D_j \) the probability \( P_i \) assigned to that disease is used as weight for the corresponding loss \( L_{rj} \).
According to the policy presented in the Introduction, we now imagine that the system is charged with the task of selecting the best treatment, i.e., the one with the least expected loss. This treatment will be denoted by $T_{\text{opt}}$, and its associated expected loss, $E_{\text{opt}}$, will be denoted by $E_i$. When expressing this procedure in mathematical terms, it is convenient to use $\{r: \text{min} \ldots\}$ as a shorthand for "the $r$ that minimizes ...:

Choose $T_{\text{opt}}$ where $r(i) = \{r: \text{min} E_r\}$. The expected loss for this preferred treatment is

$$E_i = E_{\text{opt}} = \min E_r. \quad (3)$$

The caret over $r$ serves as a reminder that $T_{\text{opt}}$ is just the system's choice; it may not be the truly best treatment to give to a patient with symptoms $x_r$.

The above reasoning is a typical application of decision theory. It presupposes that the loss scale has been suitably designed so that, for instance, a 50% risk of a loss of 2 units is neither better nor worse than an absolutely certain loss of 1 unit. Decisions are not affected if each $L_i$ is rescaled into $aL_{ij} + b$, $a$ and $b$ arbitrary constants, $a > 0$ (cf. Introduction).

So far, we have only discussed what is expected at the point in time when the system is consulted with a view to choosing a treatment. Later, the evaluator will learn what diseases the $N$ patients actually had. The actual disease of patient $H_j$ is captured in the symbol $d(i)$. To be specific, $d(i)$ is the index (= 1, 2, 3, ..., or $k$) of his actual disease. Consequently, the latter is rendered by $D_{\text{opt}}$. Thus, for instance, $d(25) = 3$ when $H_j$ suffers from $D_3$. The actual disease vector $d$ is the $N$-vector of the $r$th element of which is $d(i)$. Thus, any loss-based evaluation rests on the three bodies of data: $L$, $P$, and $d$.

When the actual disease $D_{\text{opt}}$ becomes known, it enables the evaluator to look up in the loss matrix the actual loss $L_{r(d)}$ that $H_j$ would have suffered by receiving treatment $T_r$. In particular, if the system were allowed to decide on treatment, the actual loss would be

$$\text{patient } H_j's \text{ actual loss } = L_{r(d)}, \quad (4)$$

As regards terminology, note that "actual" refers to the disease only, the treatment remaining hypothetical. To repeat, the precise designation for $L_{r(d)}$ is "the actual loss to patient $H_j$ if the system were to decide".

We take the liberty of inserting or leaving out the separating comma between subscripts in expressions like (4).

**Example B**

Consider $H_{17}$ (case No. 17 of the Acute Abdominal Pain data set), who received a probability diagnosis of $D_1$, 74.3%; $D_2$, 10.3%; and $D_3$, 15.3%; cf. Part I, Table 3. The expected loss figures associated with the four treatments of Table 1.a are:

- With $T_1$: $E_{1,17} = 0.743 + 0.103 - 0.36 + 0.153 - 47 = 10.90$, with $T_2$: $E_{2,17} = 12.86$, with $T_3$: $E_{3,17} = 10.85$, with $T_4$: $E_{4,17} = 10.48$ loss units.

Thus, despite the fact that $D_1$ is judged most probable, $T_1$ beats $T_2$ and also $T_2$ by a small margin:

$$r(17) = 4, \quad E_{17} = E_{4,17} = 10.48. \quad (B1)$$

As this patient’s actual disease was appendicitis ($D_3$), the actual loss associated with each treatment is given by Table 1, part a, column $D_3$. Thus, $T_2$ would have been the best treatment (loss $= 20$), whereas $H_{17}$’s actual loss $= L_{2(17)}d(17) = 4.42$ loss units. Therefore, $H_{17}$’s actual loss is 33–20 = 13 units in excess of what the optimal treatment would entail. In decision theory this excess is called the regret.

**Definition of regret.** The loss a person is bound to suffer even with the best treatment now available once he or she has contracted disease $D_i$ will be termed the fixed loss associated with $D_i$. It is denoted by $L_i$.

$$L_i = \min L_{ij}. \quad (5)$$

Thus, $L_2 = 20$ in the above example. Alternative terms are the inevitable or state-of-the-art loss. Each $L_i$ may be regarded as consisting of a fixed part, $L_i$, and an excess $\lambda_i$. The latter is called the regret associated with the treatment-disease combination in question. It may be viewed as the part of the patient’s loss that can be eliminated by good diagnostics. In applications that deal with treatment choice only the evaluation can be formulated entirely in terms of regret, in agreement with the fact that no system is responsible for the fixed loss.

Thus, the definition of regret and the decomposition of loss take the following form:

$$\lambda_{ij} = L_{ij} - L_i \quad \{\text{always } \geq 0\}, \quad \lambda_{ij} = L_{ij} \quad \{\text{excess}\}. \quad (6)$$

**The regret matrix $\lambda$.** (The $\lambda_{ij}$) is, therefore, obtained from the loss matrix $L = (L_{ij})$ by subtracting from each entry the smallest entry in the column, viz. $L_i$, as shown in part b of Table 1. While $L_i$’s are unconstrained, a regret matrix cannot have negative entries and there must be one or more zeros in each column. (There are two zeros in the $D_3$ column of Table 1 because $T_2$ and $T_4$ are equally optimal in case of $D_3$.)

We now have the following analogues of eqs. (1–4). The expected regret of $H_i$ when given treatment $T_r$ will be

$$e_r = \sum P_{ij} \lambda_{ij} \quad \{\text{always } \geq 0\}. \quad (7a)$$

On closer inspection it will be noted that $e_r$ and $E_r$ are the same except for an amount that does not depend on treatment (i.e., not on $r$):

$$e_r = E_r - \sum P_{ij} L_i. \quad (7b)$$
The rational choice of treatment when described in terms of regret becomes:

\[ T(i) = \text{arg min}_j \{ f(i) \} \]

Here \( T(i) \) is the same as in (2) because there is a fixed difference between \( e_i \) and \( E_{ni} \), see (7b). Moreover, the expected regret for the preferred treatment is

\[ e_i = E_{0ij} = \min_j e_{iji} \]

(8)

Finally, switching from expected to actual regret, patient \( H_i \)'s actual regret is:

\[ L_{d(0,d(0))} = L_{d(0),j} \]

(9)

For later reference we note that, as is evident from (7b) and (9), the quantity \( C_i \)

\[ C_i = \sum_j P_{ij} L_{ij} \]

(10)

plays a significant role. This is the expected fixed loss of \( H_i \) according to the system. In other words, at the point in time when the system is consulted it expects that on the average patient \( H_i \) will suffer \( C_i \) units of loss provided that, whatever disease he turns out to have, will be properly treated. Thus \( C_i \) is the system's estimate of the inherent gravity of the patient's condition or, with certain reservations, his prognosis. This difficult but important concept will be taken up again in Section 5.

**Example C**

Returning to the particular appendicitis case, \( H_{17} \), of example B, we have the following expected regret figures:

with \( T_1: e_{117} = 0.743 \cdot 0 + 0.103 \cdot 16 + 0.153 \cdot 25 = 5.47 \)

with \( T_2: e_{117} = 7.43 \)

with \( T_3: e_{117} = 5.42 \)

with \( T_4: e_{117} = 5.05 \)

Eq. (11) gives an expected fixed loss of

\[ C_{17} = 0.743 \cdot 0 + 0.103 \cdot 20 + 0.153 \cdot 22 = 5.43 \]

and it may be verified that each \( e_{ij} \) can be found by deducting this figure from \( E_{ij} \) in (B1). Once more we see that \( T(17) = 4 \), the minimal expected regret being

\[ e_{17} = e_{417} = 5.05 \]

The actual regrets associated with each treatment can be found in the \( D_i \) column of Table 1, part b. We verify that \( T_3 \) is optimal for this patient (regret = 0), but patient \( H_{17} \)'s actual regret = 5.13 loss units

as already calculated at the end of example B.

Incidentally, both the expected and the actual regrets of this patient are by far the largest amongst the 50 patients of the data set.

In the above Abdominal Pain example we also introduced actions that were not pure treatments insofar as they were aimed at gathering more information before eventually instituting a treatment. In fact, \( T_3 \) is the only alternative that can be purely therapeutic. However, regimens that involve information gathering are also covered by the above account, provided the fundamental assumptions stated at the beginning of Section 2 are met: patients with \( D_i \) should be homogeneous with respect to expected response to therapeutic and diagnostic interventions. It would take us too far to formalize this homogeneity requirement. Regimens which imply a sequence of clinical decisions usually require an in-depth analysis which will often be formulated in terms of a decision tree. For elementary accounts of this technique, see [21, 33, 36].

### 3. Data, Notation, and Terminology

The present series employs one common set of illustrative data which is derived from the Copenhagen Acute Abdominal Pain Study [4, 5]. Some key figures are given in Table 2. For details, see Part I.

**Performance measures** and certain useful related quantities will be denoted by \( Q_{51}, Q_{52}, \ldots \); \( Q \) with a subscript below 50 will refer to those defined in Parts I—III. Each performance statistic is an estimator of the score which the system would obtain if the entire hypothetical patient population were available as evaluation sample.

### Table 2: Acute Abdominal Pain data. Some key quantities used in later computations.

<table>
<thead>
<tr>
<th>Size of evaluation sample</th>
<th>Number of disease alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N = 50 )</td>
<td>( k = 3 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Evaluation sample figures employed here</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-specific</td>
<td></td>
</tr>
<tr>
<td>Abdominal Pain</td>
<td>( n_1 = 13 ), ( a_1 = 0.34 )</td>
</tr>
<tr>
<td>Appendicitis</td>
<td>( n_2 = 8 ), ( a_2 = 0.20 )</td>
</tr>
<tr>
<td>Other Diseases</td>
<td>( n_3 = 29 ), ( a_3 = 0.46 )</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
</tr>
<tr>
<td><strong>Loss</strong></td>
<td>1.00</td>
</tr>
</tbody>
</table>

In connection with each statistic we shall give the minimum and maximum possible values and two further reference values, viz. the values that two hypothetical *uninformative systems* would attain. The first of these constantly makes the non-committal assertion that all diseases are equally likely: \( P_i = k^2 \) throughout (the indifferent system). The second uninformative system quotes the relative incidence rates of the diseases but still disregards all symptoms: \( P_i = a_i \) throughout (the incidence system). Table 2 gives the values of \( k \) and \( a_i \). Consequently, these systems invariably suggest treatment \( T_2 \), as calculated in Table 3.

### Table 3: Expected regrets and suggested treatment for the two uninformative reference systems. Based on Tables 1 and 2.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Indifferent System</th>
<th>Incidence system</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>13.67</td>
<td>14.70</td>
</tr>
<tr>
<td>( T_2 )</td>
<td>3.33</td>
<td>3.40</td>
</tr>
<tr>
<td>( T_3 )</td>
<td>10.33</td>
<td>11.42</td>
</tr>
<tr>
<td>( T_4 )</td>
<td>6.00</td>
<td>4.30</td>
</tr>
</tbody>
</table>

**Suggested treatment**

\( T_2 \)  \( T_2 \)

Example: 14.70 = 0.34 · 0 + 0.20 · 16 + 0.46 · 25.
As in Parts II and III, an asterisk on the symbols $Pr_i$, $E$, and $Var$ will denote probability, expectation and variance according to the symptom-conditional probability assessments $P_{i|j}$ made by the system. Notably, $P_{ij}$ is nothing but a shorthand for $Pr_{j}(i = j)$. Many quantities in Section 2. are in fact $E^*$ expectations. For instance, $E_{ii}$ in eq. (1) is the $E^*$ expectation of the variable $L_{i|j}$ which represents the loss incurred by patient $H_i$ when receiving treatment $T_j$:

$$E^*(L_{i|j}) = \sum_{j=1}^{K} L_{ij} \cdot Pr^*(d(i) = j) = \sum_{j=1}^{K} L_{ij} \cdot P_{ij} = E_{ii}$$

(12)

cf. the wording in connection with eq. (1). All other $E_i$, $e_j$, and $C_j$ quantities can be given an analogous interpretation.

Each quality statistic $Q$ which reflects agreement between predictions $P_{ij}$ and outcomes $d(i)$ is contrasted with its $E^*$ expectation, $E^*(Q)$. The extent to which $Q$ deviates from $E^*(Q)$, denoted by $dev^*(Q)$ and calculated as $dev^*(Q) = Q - E^*(Q)$, serves as a measure of non-reliability or lack of trustworthiness. The deviation can be subjected to a significance test to see if it can be ascribed to sheer bad luck (Section 4.2. and Appendix of Part II). Some applications of this technique have been placed in Appendix B.

In particular, when $Q$ measures discriminability, $E^*(Q)$ will be a measure of sharpness, a diagnostic prediction being very sharp if one disease is declared nearly 100% certain. The resulting $dev^*(Q)$ is taken as a measure of overconfidence or diffidence respectively, as the system over- or underrates its own discriminatory ability.

One final remark concerning notation: we shall employ the shorthand

$$\text{avg}(\ldots) = \frac{1}{N} \sum_{i=1}^{N} (\ldots)$$

(13)

when dealing with the average of the expression $(\ldots)$ over patients.

4. Measures of Treatment Selection Skill

Example D

The natural first step when analysing treatment selection in terms of loss and regret is to subdivide the 50 patients of the evaluation sample according to actual disease and recommended treatment. The regret values are taken from Table 1 and the probability diagnoses and actual diseases from Part I, Table 3. The result is shown in Table 4. The simple calculations in that table allow the evaluator to judge which diseases are not handled quite so well as the others ($D_i$ in this case, the average regret amongst $D_i$, cases being no less than 3.85 units), and also to judge which treatments appear to be unduly favoured by the system ($T_j$ in this case, untimely adoption of $T_j$ being in fact responsible for the entire regret load). The figures should be compared with the uninformative systems which have $T_j$ as their favourite treatment. This implies no regret for $D_i$ and $D_j$ patients, but a regret of 10 for each of the 13 $D_i$-patients, and an average regret of 130/50 = 2.60. The regret histogram (Fig. 1) spells it out even more clearly. (In applica-

![Fig. 1: The regret analysis of Table 4 in histogram form. It is clearly seen that the entire regret load stems from X-ray examinations (T4), which would be carried out in 10 cases on non-specific abdominal pain (D3), and which would also delay laparotomy in one case of appendicitis (D2). The latter patient is No. 17, who once more stands out as the worst diagnostic miss of the sample. It is gratifying to see that T1 and T2 contribute no regret, i.e., there would be neither missed nor untimely laparotomies provided the 11 X-ray cases were properly treated in the end.]

Table 4: Regret analysis of the 50 Acute Abdominal Pain cases. The cases are subdivided according to actual disease ($D_{i|j}$) and suggested treatment ($T_{i|j}$). Within each disease-treatment entry the number of cases (1) is multiplied by the per case regret (2), taken from Table 1, to obtain an aggregate regret (3), which is summed horizontally and vertically. This in turn enables calculation of average regrets within treatments or diseases as well as overall. E.g., 50/13 = 3.85 units per case for $D_i$. The overall average regret of 1.26 units per case appears again as $Q_{64}$ in Table 6.

<table>
<thead>
<tr>
<th>Treatment suggested by the system</th>
<th>Actual disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_1$</td>
<td>$D_2$</td>
</tr>
<tr>
<td>$T_1$</td>
<td>3x0</td>
<td>0</td>
</tr>
<tr>
<td>$T_2$</td>
<td>0x0</td>
<td>10</td>
</tr>
<tr>
<td>$T_3$</td>
<td>0x2</td>
<td>0</td>
</tr>
<tr>
<td>$T_4$</td>
<td>10x5</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>3.85</td>
</tr>
</tbody>
</table>

For personal or educational use only. No other uses without permission. All rights reserved.
tions with at most 3 diagnostic alternatives it is also possible to draw a diagram that explains the actual individual regrets in terms of the assigned probabilities (see Appendix A.) Patient No. 17 suffered the largest regret: \( \lambda_{98} = 13 \), but it is a consolation to see that it was in fact he who also had the largest expected regret (compare Fig. 1 and Example C). In clinician's terms, the worst treatment choice befell the patient whose diagnosis was in fact judged least certain in the first place.

As may be expected, the performance measures in this section will all be based on loss or regret. To be more specific, each loss-based statistic has a regret-based analogue, and vice versa. Our policy will be to develop the loss-based statistics thoroughly and present the regret-based ones more briefly. The former comprise \( Q_9 \) through \( Q_60 \) but will not be presented in that order. The logic behind the indexing system will, however, soon become clear.

The loss-based measure of discriminatory ability is the average actual loss, by which we mean the average per patient loss that would be incurred if the system were to decide on treatment. It is the average value of formula (4):

\[
Q_9 = \text{avg}(L_{i|0|i}),
\]

(14)
equal to 17.22 units. This statistic and those that follow can be found in Table 5. For those readers who want to check the figure 17.22, please set up the analogue of Table 4 with loss instead of regret; the grand total becomes \( 50 \times 17.22 = 861 \).

Table 5: Acute Abdominal Pain data. The loss-based statistics for the patient sample and system under study. The layout of the table reflects the \( \text{min vs. max} \) mirror property.

<table>
<thead>
<tr>
<th>MIN</th>
<th>MAX</th>
<th>Dependent on</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( Q_{31} )</td>
<td>( L )</td>
</tr>
<tr>
<td>15.96</td>
<td>( Q_{32} )</td>
<td>( L_i, d )</td>
</tr>
<tr>
<td>15.29</td>
<td>( Q_{33} )</td>
<td>( L_i, p )</td>
</tr>
<tr>
<td>17.22</td>
<td>( Q_{34} )</td>
<td>( L_i, p, d )</td>
</tr>
<tr>
<td>17.08</td>
<td>( Q_{35} )</td>
<td>( L, p )</td>
</tr>
</tbody>
</table>

The sharpness statistic associated with the average actual loss is the average loss the system would expect the patient to incur if its therapeutic advice were followed (i.e., assuming perfect trustworthiness of the probabilities \( P_j \)). To the physician it is the loss to be expected when the system's advice can be trusted. This average expected loss amounts to

\[
Q_{60} = E^o(Q_{60}) = \text{avg} \ \text{min} \ \text{E}_r = \text{avg} \ \text{E}_n
\]

(15)

\( E_r \) and \( E_n \) being defined in (1—3), cf. (12). If \( Q_{60} \) is (significantly) lower than \( Q_{9} \), we conclude that the system overestimates its own cleverness at treatment choice (overconfidence).

For a specified group of patients, the average actual loss is bounded on both sides. The upper bound is achieved when every patient is given what is for him or her the worst possible treatment: \( Q_{9} = \text{avg} \ \text{max} \ L_{i|0|i} \).

Similarly, the average actual loss per patient assuming optimal treatment or, in other words, the average fixed loss becomes a lower bound:

\[
Q_{60} = \text{avg} \ \text{min} \ L_{i|0|i} = \text{avg} \ L_{i|0|i},
\]

(17)
cf. the definition of the fixed loss \( L_i \). (5)

Again, if the system were asked to provide estimates of \( Q_{60} \) and \( Q_{9} \) without knowing the actual diseases of the patients, it would calculate the expectation of both quantities:

\[
Q_{60} = E^o(Q_{60}) = \text{avg} \ \sum_i P_i \ \text{max} \ L_{ij},
\]

(18)

\[
Q_{9} = E^o(Q_{9}) = \text{avg} \ \sum_i P_i \ \text{min} \ L_{ij} = \text{avg} \ C_i
\]

(19)

The latter quantity being the average expected fixed loss; \( C_i \) the expected fixed loss for the \( i \)-th patient, was defined in (11).

The quantities \( Q_{60} \) and \( Q_{9} \) reflect the composition of the sample of patients at hand, and these quantities again find their bounds in some quantities that are based on the loss matrix alone. \( Q_{9} \) and \( Q_{60} \) have an upper bound of

\[
Q_{60} = \text{max} \ L_i
\]

(20)

and a lower bound of \( \text{min} \ L_i \). \( Q_{60} \) and \( Q_{9} \) have a lower bound of

\[
Q_{9} = \text{min} \ L_i
\]

(21)

and an upper bound of \( \text{max} \ L_i \).

The system of the \( Q \)-statistics will now be explained in so far as it has not yet become clear. There is a twin-structure in the \( Q \)-measures: \( Q_{31} \sim Q_{60}, Q_{32} \sim Q_{61}, Q_{33} \sim Q_{62} \) and \( Q_{34} \sim Q_{63} \) are all twin pairs. The high-indexed \( Q \)'s have to do with large loss-based quantities, \( \text{max} \) appearing in the formulae. The low-indexed \( Q \)'s, on the contrary, have to do with small quantities: \( \text{min} \) appears instead of \( \text{max} \). The sequences \( Q_{31} \) through \( Q_{34} \) and \( Q_{60} \) through \( Q_{63} \) are mirror images of each other, with the mirror placed between \( Q_{31} \) and \( Q_{60} \) (the layout of Table 5 reflects this mirror structure).

Thus, it is now easy for us to complete the picture by supplying two more \( Q \)'s: \( Q_{64} \), the mirror image of the average actual loss \( Q_{9} \), is the average actual loss when the treatment which is judged worst by the system is decided on:

\[
Q_{9} = \text{avg} \ L_{i|0|i},
\]

(22)

with \( r_{ij} = (r: \text{max} E_r); \) and \( Q_{64} \) (image of \( Q_{9} \)), is the system's expectation of \( Q_{64} \):

\[
Q_{64} = E^e(Q_{64}) = \text{avg} \ \text{max} \ E_r
\]

(23)

\( Q_{64} \) is high when the system is good at identifying treatments that are to be avoided by all means. If \( Q_{64} \) exceeds \( Q_{9} \), it means that the system is overconfident regarding its own ability to do so.

This completes the description of the ten loss-based measures. Note that in Table 5 we made explicit the dependence of the various measures on our three data sources: \( L \) (agreed loss matrix), \( P \) (output from system), and \( d \) (diagnostic work-up).

The scores achieved by the system under study in our Acute Abdominal Pain example are given in that table. The figures suggest, as is in fact readily verified from formulae (14) up to (23), that the following inequalities hold:

\[
Q_{31} \leq Q_{32} \leq Q_{33} \leq Q_{34} \leq Q_{60},
\]

(24)

\[
Q_{51} \leq Q_{52} \leq Q_{53} \leq Q_{54} \leq Q_{64},
\]

and \( Q_{9} \) is also larger than \( Q_{64} \), unless the system's preferred treatments are on the average worse than the treatments which are judged most harmful by the system — which is indeed unlikely to occur.

The whole story up to this point can be repeated for regret, and we thus derive ten regret-based statistics which are numbered \( Q_{9} \) through \( Q_{60} \) where \( Q_{9} \) is the regret analogue of \( Q_{9} \), etc. The only change in the formulae is the replacement of the loss \( L_{ij} \) by the regret \( R_{ij} = L_{ij} - L_i \) and \( E_r \) by \( E_s \), cf. Section 2. The inequalities (24) also hold for the regret statistics, but by the very definition of regret the specific equality \( Q_{9} = Q_{60} = Q_{54} = 0 \) holds. Moreover there are simple relations between regret and loss statistics:

\[
Q_{58} \sim Q_{65} = Q_{56} = Q_{64} = Q_{58} = Q_{54} = Q_{53},
\]

(25)
Table 6: Acute Abdominal Pain data. Scores by the actual system, the indifferent system, and the incidence system on all 20 loss- and regret-based statistics.

<table>
<thead>
<tr>
<th>LOSS</th>
<th>REGRET</th>
<th>Depend on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual system</td>
<td>Indiff. system</td>
<td>Incid. system</td>
</tr>
<tr>
<td>Q31</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Q32</td>
<td>15.96</td>
<td>15.96</td>
</tr>
<tr>
<td>Q33</td>
<td>15.29</td>
<td>14.00</td>
</tr>
<tr>
<td>Q25</td>
<td>17.22</td>
<td>18.56</td>
</tr>
<tr>
<td>Q28</td>
<td>17.08</td>
<td>17.33</td>
</tr>
</tbody>
</table>

Table 6 gives the values for all twenty statistics including, for reference purposes, those of the indifferent and the incidence systems. It is seen that the system would have caused the patients an average regret of 1.26 units (Q64). This figure is roughly equivalent to treating 7 out of 8 patients correctly and subjecting the eighth to a low-risk laparotomy (10 units of regret, Example A). The two uninformative systems both recommend immediate laparotomy (Table 3). This would have implied an average regret of 2.60 units (= Q64); so by this standard the system would be able to roughly halve regret. By way of comparison, our system had an error rate of 20% while that of the incidence system was 42% (Part III, Table 2). The same comparison in terms of loss tells us that the average loss incurred when using the actual systems is only about 8% less than when using an uninformative system (17.22 versus 18.56). Thus the ratio of 2.60 - 1.26 = 1.34 units is a large fraction of the loss that the average loss incurred when using the actual system regret, but it is small compared with the total misery involved in admissions for acute abdominal catastrophes. (The interpretation of such ratios between amounts of total misery depends on what is regarded as baseline misery (zero loss), here pains that subside, a quick discharge from hospital, and absence of sequelae; this is true of any relative statement concerning loss, as opposed to regret comparisons where a natural zero exists.)

The large range of values that the average loss may take for the patient sample at hand, from 15.96 to 35.62, is estimated by the system to run from 15.29 to 34.34. This underestimation is explained by the most serious «disease» (D3) being underassigned (Part II). The actual loss of 17.22 was expected by the system to be very slightly less, 17.08, but the regret of 1.26 was expected to be nearly 50% larger, Q65 being 1.80. This is a divergent trait that we have noted before. The pattern \( Q_{64} < Q_{65} < Q_{66} < Q_{67} > Q_{68} > Q_{69} \) is typical of difference.

The question remains whether or not the deviations from perfect reliability can be explained by chance alone. This problem is addressed in Appendix B.

5. Prognosis

5.1. The Role of Fixed Loss in Prognostics

The authors have tried to steer clear of the ambiguities of the term diagnosis. For instance, we speak of a patient's actual disease and not his actual diagnosis. The term prognosis has even more facets of meaning. Although a prognostic statement is necessarily a prediction conditional on medical action, it is often not made explicit on what action. The statement may, for instance, be conditional on a specified treatment regimen, or on optimal treatment, or just on treatment of standard quality. The term «spontaneous prognosis» often refers to the course of events to be expected when curative («causal») therapy is withheld, not when the patient is literally denied care.

Which came first — the egg or the hen? We are reminded of this old riddle when we ask: does treatment choice determine prognosis or does prognosis determine the choice of treatment? However, as soon as we realize the multiple meanings of «prognosis», the paradox disappears. The sequence of events to be expected when curative («causal») therapy is withheld, not when the patient is literally denied care.

In either case the problem of treatment choice vanishes: in the first case only one regimen is involved; in the second there will eventually be enough information to ensure that the choice becomes trivial. Consequently, all regrets become zero. And in either case \( D_1, \ldots, D_6 \) may be of two kinds:

a. They may be treatment-prognostic units as in the preceding sections. According to eq. (5), the fixed loss \( L_i \) is the loss associated with the course of events when \( D_i \) is present and appropriately treated.

b. Or else the classes \( D_1, \ldots, D_6 \) may stand for different courses of illness such as «12—15 months of survival with major sequelae», and each \( L_i \) will then stand for the loss associated with the corresponding course, \( D_i \).
It is essential to note the distinction between two kinds of clinical skill: one has to do with regret and decision on action; the other has to do with prognostication, where fixed loss plays a prominent role. Imagine, for example, a stream of patients with appendicitis, salpingitis, bowel obstruction due to malignancy, or carcinosis of the peritoneum. One diagnostian (man or machine) may be good at selecting the right patients for immediate operation (appendicitis and bowel obstruction), another at identifying patients with a gloomy long-term prognosis (the last two categories). So we do need two separate sets of performance assessment tools: one for analysis of regret, and another which compares predicted and actual fixed loss. The authors consider this an important point.

In order to discuss how to measure the quality of a prognostic statement, we will consider a particular patient: The probability diagnosis of one of the Copenhagen patients (viz. \( H_{13} \)) was as follows: \( D_1, \) 64\%; \( D_2, \) 12\%; \( D_3, \) 24\%. Effectively, then, the system predicted that with appropriate treatment this patient would suffer no loss with probability 64\%, 20 units with probability 12\%, and 22 units otherwise; cf. \( L_1, L_2, L_3 \) in Table 1. This probabilistic prognosis conditional on appropriate treatment is shown in cumulative distribution form in Fig. 2. Its mean, which predicts how bad the patient is off "on the average," is nothing but the quantity \( C_i \), eq. (11): 

\[
C_{13} = 0.64 \times 0 + 0.12 \times 20 + 0.24 \times 22 = 7.68 \text{ loss units. (26)}
\]

This figure is perhaps not so easy to interpret in our clinical context; instead, the reader may prefer to think of cumulative proportion, i.e.: 

\[
\Pr'f\{Ld(i) \leq z\} = \frac{1}{N} \sum_{i=1}^{N} I(Ld(i) \leq z).
\]

Now since this patient, \( H_{13} \), proved to have \( D_3 \) (d(13) = 1), the actual fixed loss was \( L_{d(13)} = L_1 = 0 \). Somehow we must find ways of penalizing the discrepancy between the distribution in Fig. 2 and the observed value. Presumably, the first idea that comes to the reader’s mind is to calculate the following squared deviation:

\[
(L_{d(i)} - C_i)^2 = (0 - 7.68)^2 = 58.98 \text{ (loss units)}^2. \quad (27)
\]

The underlying idea would be that, quite apart from your disability and suffering, you incur a special loss by being misinformed about your future, and this special loss might well be taken as proportional to the squared distance between the predicted mean amount of suffering and the actual amount.

5.2. Measures of Prognostic Skill

The natural point of departure is Fig. 2 which looks in the general case as shown in Fig. 3a. When this diagram is drawn for each patient, we get \( N \) diagrams that jointly contain the requisite information. Technically speaking, the diagram shows the predicted and the observed cumulative distributions of the quantity \( L_{d0} \) which describes the fate of patient \( H_i \). The precise interpretation of \( L_{d0} \) may vary as indicated above. Note that the equations below do not presuppose that \( L_1, \ldots, L_N \) are in increasing order although this has been assumed in the figure for ease of presentation.

The observed distribution function is a one-step function because there is only one observation: the ordinate jumps from zero to one for abscissa \( z \) equal to the observed \( L_{d0} \). The predicted distribution curve displays the probability assigned to outcomes no worse than \( z \) loss units, denoted here by \( G(z) \), i.e.:

\[
G(z) = \Pr'f\{L_{d0} \leq z\}. \quad (28)
\]

Using the notation \( f(\text{true}) = 1, f(\text{false}) = 0 \), this definition may also be written \( G(z) = \sum_{j} I(L_j \leq z) P_j \). It is expedient to have a brief notation for the expectation and the standard deviation of this predicted distribution. The expectation \( C_i \) was already defined in (11):

\[
C_i = \sum_j L_j P_j = E[z \{L_{d(i)}\}],
\]

cf. (26). The variance is

\[
s^2 = \sum_j (L_j - C_i)^2 P_j = \text{Var}[z \{L_{d(i)}\}],
\]

and the standard deviation \( s_i \) is its square root. Thus prepared, we are ready to define a few statistics that throw light on prognostic performance.
Some of the statistics that were introduced in Section 4. already tell us something about this question. We have computed the average fixed loss:

\[ Q_{72} = \text{avg} \ L_{d(i)} = 15.96 \text{ loss units} \]  

(31)

and compared it with

\[ Q_{53} = \text{avg} \ C_1 = 15.29. \]  

(32)

We find the resulting reliability statistic,

\[ Q_{71} = Q_{53} - Q_{53} = \text{dev}^* \{Q_{53}\} = +0.67, \]  

(33)

to be on the positive side, indicating that the prognostications as a whole were a bit optimistic (though only 0.67 loss units per patient).

Unlike these statistics which look for a systematic trend the measures of misprognostication that follow examine the individual predictions and penalize disagreement with observed outcomes.

A squared deviation measure was suggested at the end of Section 5.1.:

\[ Q_{72} = \text{avg} \ (L_{d(i)} - C_1)^2 = 54.01 \text{ (loss units)}^2. \]  

(34)

This figure and the corresponding scores for the uninformative systems, which are nearly twice as large, are given in Table 7. The probabilistic predictions themselves lead us to expect the following figure:

\[ Q_{73} = E^\circ \{Q_{72}\} = \text{avg} \ E^\circ \{(L_{d(i)} - C_1)^2\} = \text{avg} \ \sigma^2 = 65.05. \]  

(35)

The deviation

\[ Q_{74} = \text{dev}^* \{Q_{72}\} = Q_{72} - Q_{73} = -11.04 \]  

(36)

is negative indicating that the system expected itself to perform worse than it actually did (a diffident trait).

It is intuitively appealing to use the following quantity (call it \( q^2 \)) as a measure of the extent to which the predicted distribution fails to cluster around the observed value, \( L_{d(i)} \):

\[ q^2 = \sum_j (L_j - L_{d(i)})^2 P_j \]  

(37)

On closer analysis it turns out that in order to obtain a scoring rule with reasonable properties the individual \( q^2 \)'s must be weighted by the reciprocal of \( s_i \). Hence we get

\[ Q_{75} = \text{avg} \{q^2/s_i\} \text{ (loss units)} \]  

(38)

After some algebra one obtains an alternative formula:

\[ Q_{73} = \text{avg} \ g((L_{d(i)} - C_1)^2/s_i + s_i). \]  

(39)

The associated \( E^\circ \) expectation is

\[ Q_{76} = E^\circ \{Q_{71}\} = \text{avg} \{2s_i\}, \]  

(40)

and the reliability statistic becomes

\[ Q_{77} = Q_{75} - Q_{76} = \text{dev}^* \{Q_{73}\}. \]  

(41)

Being negative like \( Q_{72} \), \( Q_{77} \) also suggests diffidence (Table 7). However, the diffident trait is not statistically significant; for instance, \( Q_{74} \) is only \(-0.86\) times the relevant standard deviation as calculated in Appendix B.

As indicated in the formulae above, a linear transformation of the loss scale \((L_i \rightarrow aL_i + \beta, a > 0)\) will multiply \( Q_{72} \) by \( a^2, Q_{76} \) by \( a \).

It should be obvious that the techniques of this section can be generalized to cover situations in which patient outcomes are measured on a continuous loss scale (Fig. 3b). The various sums over \( j \) now become integrals (with respect to \( dG(x) \)). There exists a bewildering number of scoring rules for misprognostication, applicable to the discrete and the continuous cases alike; (see, e.g., [27]). But we feel that \( Q_{73} \) and \( Q_{75} \) will be among the ones that are easiest to calculate and interpret, yet guaranteed to detect poor prognostication in a fairly efficient manner.

We have come to question his competence. If we could measure loss faith in the physician's cheerful attitude and perhaps lose faith in the physician's cheerful attitude and perhaps come to question his competence. If we could measure loss faith in the physician's cheerful attitude and perhaps lose faith in the physician's cheerful attitude and perhaps come to question his competence. If we could measure loss faith in the physician's cheerful attitude and perhaps lose faith in the physician's cheerful attitude and perhaps come to question his competence. If we could measure loss faith in the physician's cheerful attitude and perhaps lose faith in the physician's cheerful attitude and perhaps come to question his competence.

We have fallen back on this purely mathematical criterion for lack of a more cogent medical or psychological criterion of what constitutes a good prognostication or prognosticator. We should like to have a theory at our disposal which would quantify the displeasure associated with illness — an apprehension which may be mitigated by a com-

Table 7: Acute Abdominal Pain data: two measures of misprognostication and the associated reliability statistics. Each measure (top) deviates from its \( E^\circ \) expectation (middle) by an amount (bottom) which constitutes a measure of non-reliability. In brackets are shown the minimum possible value of the statistic, the score of the indifferent system, that of the incidence system and, finally, the maximum possible value; these four reference values generally depend on the figures of Table 2 as well as on the fixed losses (Table 1).

<table>
<thead>
<tr>
<th>Misprognostication statistic (low values indicate good performance)</th>
<th>( Q_{72} = 54.01 )</th>
<th>( Q_{75} = 13.23 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E^\circ {\text{Misprog.}} ) (low values when predictions are sharp)</td>
<td>( Q_{73} = 65.05 )</td>
<td>( Q_{76} = 15.23 )</td>
</tr>
<tr>
<td>Reliability statistic, ( \text{dev}^* {\text{Misprog.}} ) (near-zero values indicate reliable predictions)</td>
<td>( Q_{74} = -11.04 )</td>
<td>( Q_{77} = -2.00 )</td>
</tr>
</tbody>
</table>

Note: \( Q_{74} = -80.5 \) is attained when \( d(i) = 2 \) and \((P_{i1}, P_{i2}, P_{i3}) = (13/44, 0, 31/44)\). Minimum of \( Q_{77} \): not calculated.
for distinguishing good prognostic statements from poor ones.

Before closing this section, let us mention two related techniques. Firstly, a fractile-based reliability graph was defined in Part II (Discussion). This graph supplements $Q_n$. Related techniques can be found in [32].

Secondly, it may happen that the $k$ prognostic alternatives can be ranked according to severity but not quantified. For instance, the system evaluators may agree that $L_1 < L_2 < \ldots < L_5$, without wishing to assign numbers to these $L_i$'s. In other words, the horizontal scale of Fig. 3.a is now no longer an interval scale but only an ordinal one. If so, the evaluation must be based on comparing the ordinates of the two step curves in each of the intervals between $L_i$'s. One of the best-known loss matrix distributions is the so-called ranked quadratic scoring rule which will be briefly explained in Appendix C.

6. Variable Loss Figures

6.1. Uncertainty within Clinical Decision Tasks

In the main part of the present Section 6, we shall look into the problems connected with variation of the loss figures from one clinical decision task to the next (i.e., between-encounter variability). But it is essential first to clarify the three main sources of uncertainty and variability within a given decision task (within-encounter loss variability):

1. The measurement problem: Consequently, with an accurate description of the patient's preferences and a very accurate diary of what happened in the course of his illness there is no accurate way of determining the associated subjective loss.

2. Uncertainty about the course of illness: what the diary will contain cannot be predicted with certainty, even if the doctor has complete knowledge of the patient's disease, of his condition at the moment of clinical decision (e.g. his operative risk status), and of the treatment plan.

3. Uncertainty about patient's preferences: for various reasons the typical patient-doctor encounter does not involve much probing and sometimes even does not allow probing into the individual patient's attitudes and wishes.

It is quite legitimate to imagine the clinical decision-maker handling these three factors by mental averaging over a personal distribution which reflects his uncertainty, the result being a set of loss figures with the properties discussed in Section 2, if $D_0$ is present and $T_1$ is given, then in view of the available data the mean loss will amount to a certain number, and this becomes the decision-maker's $L_0$. This is the way in which the «agreed loss figures» of Section 2. are thought to arise.

6.2. Variability between Clinical Decision Tasks

Between-encounter loss variability is something that any physician is familiar with. Just think of two identical wrist fractures that have to be treated differently because the patient's occupations differ. In utility terms, they attach different loss values to some of the possible outcomes of treatment.

This (fourth) level of loss variability acquires a special significance in the context of evaluation of a system which is to serve as a decision aid because the individual loss matrices of the clients will usually be unknown to the system staff. When faced with this fact, most people would agree that those who are to market the system must prove that the device has a good mean performance in terms of the mean for all clinical encounters. Consequently, the evaluator must view the figures in the loss matrix as stochastic variables. Their distribution should reflect the variation of the loss matrix from one clinical encounter to the next, as assessed in an empirical pilot study or agreed upon by expert consensus.

**Example E**

Imagine two diseases $D_1$ and $D_2$. Let $T_1$ and $T_2$ be the treatments appropriate for $D_1$ and $D_2$, respectively. Let $T_3$ stand for a risky but definitive diagnostic test followed by $T_1$ or $T_2$ as appropriate.

We shall illustrate the implications of between-encounter loss variability by the simplest possible instance: Only two easily distinguished classes of patients are envisaged, each class having its own regret matrix (Table 8). As is clear from these two radically different

<table>
<thead>
<tr>
<th>Regret matrices</th>
<th>“Class I”</th>
<th>“Class II”</th>
<th>Mean of I and II</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>0 96</td>
<td>0 4</td>
<td>0 50</td>
</tr>
<tr>
<td>$T_2$</td>
<td>4 0</td>
<td>96 0</td>
<td>50 0</td>
</tr>
<tr>
<td>$T_3$</td>
<td>2 2</td>
<td>50 50</td>
<td>26 26</td>
</tr>
</tbody>
</table>

matrices, $T_1$ will hardly ever be chosen for a class I case, whereas it is usually preferred in class II cases. (Exceptions occur when the diagnostic odds are extreme.) In each particular case, the rational decision only depends on the regret matrix of the appropriate class. The doctor does not allow that of the other class to influence his decision. In particular, even if we assume that the two classes occur with exactly the same frequency, the regret values of an imaginary «mean class» shown on the right will not be of any use. For instance, $T_3$ would never be a worthwhile option, judging from the «mean matrix», despite the fact that it is the action of choice in many class I cases.

If a country-wide system is consulted by doctors over telephone lines, its value to the community must be calculated as a mean benefit over the two classes of patients. The result will in general be different from what one would get by ignoring inter-class variation and using the mean matrix instead. The analyses of Section 4, which used the same regret matrix for all 50 patients, can rightly be criticized for not taking this aspect into account. Summing up, it is the mean consequence across loss configurations that counts, not the consequence of any mean configuration.

6.3. A Regret-based Quality Measure when Loss Matrices are Stochastic

This additional step of forming a mean over several or an infinite ensemble of loss configurations is what distinguishes applications in which one wants to take between-encounter variability into account, from those which involve only within-encounter variability and for which the methods of Section 4, are sufficient (cf. the last remarks in Section 6.1). The actual loss of a particular evaluation patient must be replaced by a mean value calculated over a distribution of loss matrices. This in turn implies that the average over patients, $Q_{mean}$, is replaced by a doubly averaged actual loss, viz. across $L$-configurations as well as across patients. We speak of variable-loss/regret performance scores.

Quality measures of this kind tend to involve complicated mathematics because of the need to integrate over some suitably chosen loss matrix distribution. However, as we are anxious to illustrate the approach and have promised to offer
a well-filled kit of evaluation tools, we shall as a compromise
present one such quality measure, \( Q_{\theta} \).

After a brief outline of the mathematical assumptions we
shall leave it to the graphs and the numerical example to
provide the potential user with an intuitive appreciation of
the approach.

The instance we shall consider concerns regret evaluation
in the following two-disease three-action situation, similar to the one in Example E:

\[
\text{Regrets} \quad \begin{array}{c|c|c}
T_1 & D_1 & D_2 \\
K_1 & 0 & K_2 \\
K_2 & K_3 & K_3 \\
\end{array} \quad (K_1, K_2, K_3 \geq 0) \quad (42)
\]

Thus, \( T_1 \) (\( T_2 \)) is the treatment appropriate for \( D_1 \) (\( D_2 \)),
while \( K_2 \) reflects the risk and inconvenience associated with
a confirmatory test followed by appropriate treatment, \( K_1 \)
being independent of the patient’s disease. As to the variation
of the regrets, \( K_1 \), \( K_2 \), and \( K_3 \) across the population of
clinical encounters, we assume them to have a mean value of
\( m_1 \), \( m_2 \), and \( m_3 \), respectively. The only further parameter of
our model is \( \beta \) (\( \beta > 0 \)), which governs the concentration
around the mean values. The larger \( \beta \) is, the more the dis­
tribution is bell-shaped and narrowly concentrated around
the mean. If \( \beta \) is nearly 0, the distribution is highly skewed
towards large values and has a large coefficient of variation.
The three distributions are assumed to be independent and
to have the same coefficient of variation, i.e. the same
shape: Weibull.

For the sake of completeness, we must add that the shape of the
assumed distribution is of the so-called Weibull type (see,
\textit{e.g.}, \( [20] \)), but the reader is not assumed to be familiar with
this distribution. Summing up, we assume:

\[
\begin{align*}
\text{mean}\{K_1\} &= m_1, \\
\text{mean}\{K_2\} &= m_2, \\
\text{mean}\{K_3\} &= m_3, \\
K_1, K_2, K_3 &\text{ independent;} \\
\text{coefficient of variation: governed by } \beta; \\
\text{shape: Weibull.}
\end{align*}
\]

Under these assumptions, we have derived the following
variable-regret analogue of the average actual regret score
(\textit{i.e.}, of \( Q_{\theta}\)):

\[
Q_{\theta} = \text{avg} \left( \frac{(m_{10}P_{01})^{-\beta} + m_1^{-\beta}}{(m_{11}P_{11})^{-\beta} + (m_{21}P_{11})^{-\beta} + m_2^{-\beta}} \right) \
\]

The associated \( E^\beta \) expectation is:

\[
Q_{\#} = E^\beta (Q_{\theta}) = \text{avg} \left( \frac{(m_{10}P_{01})^{-\beta} + m_1^{-\beta}}{(m_{11}P_{11})^{-\beta} + (m_{21}P_{11})^{-\beta} + m_2^{-\beta}} \right)^{1/\beta} \
\]

(43)

We refer to this as the \textit{mean regret} (\textit{across patients}) \textit{mean regret}
(\textit{actual disease}) \textit{average actual regret} and \( Q_{\#} \) the \textit{average mean expected \( \textit{across diseases} \textit{regret.}

As a first illustration, we take as our mean regrets:
\( m_1 = 10, m_2 = 16, m_3 = 2 \).

\text{Fig. 4 gives} \( Q_{\theta} \) and \( Q_{\#} \) for a single patient
as a function of the probability assigned to \( D_1 \) (\textit{i.e.},
\( P_{11} \)) for \( \beta = 1, 2, 8, \text{and } \infty \). The value \( \infty \) brings us back to the
\textit{deterministic} situation of the previous sections where the
regret is fixed at 0, 2, 10, or 16 as appropriate. As we move
towards lower \( \beta \) values, the effect of regret variability be­
comes more and more pronounced.

Next we examine the effect of varying the mean price of
a safe diagnosis, \( m_3 \). \text{Fig. 5 examines the values} \( m_3 = 0, 2, 4,\)
10, and \( \infty \). When \( m_3 \) is zero, all regret is eliminated and all
the curves coincide, so we have added the small arrows to
guide the reader’s eye.

Example F

Consider the three diagnostic categories of the Acute
Abdominal Pain study. Assume that by some magic
trick all cases of \( D_3 \) have been filtered away so that only
\( D_1 \) and \( D_2 \) have to be distinguished. Imagine also that in
addition to $T_1$ (wait and see) and $T_2$ (immediate operation) we have at our disposal a completely reliable diagnostic test $T_3$ (perhaps laparoscopy), with mean regret $m_3 = 2$ for both diseases. This brings us back to the situation we have just considered, provided the regret figures 10 and 16 in Table 1 are regarded as the means $m_1$ and $m_2$ of a stochastic regret distribution. Let us take the assumptions underlying (44–45) for granted, $\beta$ being 2. Applying these formulas to the 21 patients with $D_1$ or $D_2$ yields Table 9. (The assigned probabilities have been rescaled to reflect the assumption that $D_1$ has been eliminated.) In order to allow the reader to appreciate the behaviour of the scoring rule, the table shows each individual contribution and also what that contribution would have been if the actual disease had proved to be the opposite. The individual $Q_{st}$ terms are also shown. Compared with the two reference systems, our system entails only one third as much regret the main part of which is as usual due to case No. 17.

The relationship between the variable-regret measure $Q_{st}$ and its deterministic counterpart $Q_{d}$ is in a sense analogous to that between the error rate and the continuous scoring functions of Part III. In the two-disease situation it is natural to argue that the abrupt jump in the error rate from 0 to 1 at $P_{d_0} = 0.5$ is not in harmony with clinical use of diagnostic probabilities: continuous scoring functions should be used. The analogous grievances about $Q_{a}$ (and also $Q_{st}$) run as follows: there are exactly specifiable $P_{a_0}$ values at which a patient’s actual regret and with it $Q_{st}$ will make an inordinate jump. Consider, for example, the jumps of the deterministic ($\beta = 1$) curve in Fig. 4 when the actual disease is $D_1$; a probability $P_{d_1} = 0.1999$ means an actual regret of 10 units, and a probability of $P_{d_1} = 0.2001$ an actual regret of only 2 units. Indeed, one cannot feel more comfortable with one of the variable-regret curves of this figure. True, the curves were derived under certain assumptions concerning the statistical distribution of the regret values over clinical encounters. But they may also be viewed simply as intuitively appealing continuous scoring functions that are no more, and probably less, arbitrary than those discussed in Part III. Other distributional assumptions will give rise to qualitatively similar curves. It can safely be said that the step from deterministic to stochastic is the really big one, and exactly which type of distribution is stipulated will not matter as much as the assumed amount of variability, here governed by $\beta$.

The paper by Murphy [30] seems to be the earliest paper that recognises the potentials of variable-loss scoring methods (*expected utility scores* in his later terminology). For more recent discussions, see Murphy [31] and Pearl [35].

The present approach has even more potential flexbility. There is no need to assume that all patients have the same loss distribution. One may think of $m_1$, $m_2$, $m_3$, and $\beta$ as varying from patient to patient. Strange as it may seem, not only the assigned probabilities but also the various coefficients in (44) will now be patient-dependent, both being functions of $x_0$, i.e., of the body of data describing the rth clinical encounter. But there is nothing to prevent that, and settings may easily be imagined in which this is the rational scoring strategy.

7. Discussion

Excellent introductions to decision theory and its application in medicine can be found in the two 1975 special issues [21] and [33], and in Card [7]. As this Discussion section mainly proceeds from the general to the specific, it must be understood that these tutorials and the other texts of a general nature are not always quoted when we come to more specific issues that they also discuss. A critical note is struck by Ranshoff and Feinstein [37], Szolovits and PAUKER [41] and, in the context of clinical chemistry, by HILDEN [16]. Decision analysts often relate benefits to costs (cost/benefit, cost/effectiveness, etc.). BUNKER et al. [6] (in particular chapters 1 and 10) have a valuable discussion of these con-
Table 9: Application of variable-regret statistics to (modified) abdominal pain data, Example F. Parameters: $m_1 = 10$, $m_2 = 16$, $m_3 = 2$, $\beta = 2$. The original data can be found in Part I, Table 3, the first 21 patients. A summary of the method of computation is given in the note below the table.

<table>
<thead>
<tr>
<th>Patient No. $i$</th>
<th>Probability assigned to $D_i$ (%)</th>
<th>$d(i) = 1$</th>
<th>$d(i) = 2$</th>
<th>Mean expected regret</th>
<th>Actual disease</th>
<th>Mean actual regret</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>93.8</td>
<td>.238</td>
<td>10.718</td>
<td>.968</td>
<td>1</td>
<td>.238</td>
</tr>
<tr>
<td>2</td>
<td>98.8</td>
<td>.002</td>
<td>15.787</td>
<td>.186</td>
<td>1</td>
<td>.002</td>
</tr>
<tr>
<td>3</td>
<td>99.8</td>
<td>.000</td>
<td>15.994</td>
<td>.030</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>84</td>
<td>1.024</td>
<td>4.310</td>
<td>1.564</td>
<td>1</td>
<td>1.024</td>
</tr>
<tr>
<td>5</td>
<td>89</td>
<td>.631</td>
<td>6.782</td>
<td>1.336</td>
<td>1</td>
<td>.631</td>
</tr>
<tr>
<td>6</td>
<td>96</td>
<td>.067</td>
<td>13.609</td>
<td>.653</td>
<td>1</td>
<td>.067</td>
</tr>
<tr>
<td>7</td>
<td>86</td>
<td>.857</td>
<td>5.237</td>
<td>1.477</td>
<td>1</td>
<td>.857</td>
</tr>
<tr>
<td>8</td>
<td>98.6</td>
<td>.023</td>
<td>15.699</td>
<td>.221</td>
<td>1</td>
<td>.003</td>
</tr>
<tr>
<td>10</td>
<td>94</td>
<td>.146</td>
<td>12.078</td>
<td>.824</td>
<td>1</td>
<td>.146</td>
</tr>
<tr>
<td>11</td>
<td>94</td>
<td>.197</td>
<td>11.290</td>
<td>.909</td>
<td>1</td>
<td>.197</td>
</tr>
<tr>
<td>12</td>
<td>90</td>
<td>.459</td>
<td>8.256</td>
<td>1.203</td>
<td>1</td>
<td>.459</td>
</tr>
<tr>
<td>13</td>
<td>84</td>
<td>.985</td>
<td>4.509</td>
<td>1.546</td>
<td>1</td>
<td>.985</td>
</tr>
<tr>
<td>14</td>
<td>16</td>
<td>5.107</td>
<td>.508</td>
<td>1.256</td>
<td>2</td>
<td>.508</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>7.920</td>
<td>.118</td>
<td>.773</td>
<td>2</td>
<td>.118</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>8.642</td>
<td>.060</td>
<td>.617</td>
<td>2</td>
<td>.060</td>
</tr>
<tr>
<td>17</td>
<td>88</td>
<td>.697</td>
<td>6.292</td>
<td>1.381</td>
<td>2</td>
<td>6.292</td>
</tr>
<tr>
<td>18</td>
<td>1.0</td>
<td>9.960</td>
<td>.000</td>
<td>.103</td>
<td>2</td>
<td>.000</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>9.398</td>
<td>.017</td>
<td>.405</td>
<td>2</td>
<td>.017</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>9.471</td>
<td>.014</td>
<td>.379</td>
<td>2</td>
<td>.014</td>
</tr>
<tr>
<td>21</td>
<td>2.2</td>
<td>9.817</td>
<td>.003</td>
<td>.221</td>
<td>2</td>
<td>.003</td>
</tr>
</tbody>
</table>

Average for present system $Q_{88} = .815$ $Q_{87} = .569$

Average for incidence system 1.815 1.817
Average for indiff system 1.809 1.843

Summary of calculations: (2) = $P_{d1}(D_1 + P_d2)$, from original data; (3) & (4), cf. graph; (5) = $(2)(3) + (1 - (2))\times(4)$, cf. graph; (6) from data; (7) equal to (3) or (4) as appropriate.

cepts and many fascinating applications, including Neutra's analysis of the handling of appendicitis (Ch. 18). For a very recent bibliography, see [24].

Decision theory — like other mathematical disciplines — must undergo a translation process before being applied. The entire abstract machinery (with its states, actions, observations, utility, etc.) has to be fitted to medical entities (diseases, treatments, symptoms, human benefits, etc.), and this is by no means an easy process (see for example the short but lucid overview by Card [8] and his paper [7]). The man-headed monster that goes under the deceptively simple name of 'disease' is perhaps the worst trouble maker. Komaroff [23] explains this very well in a paper which also, more generally, discusses the softness of most medical data in terms intended for a technical audience (also see [25]). Of particular importance is Card's [8] dismissal of the suggestion that one might proceed directly from symptoms to treatment, thus bypassing the stepping stone of 'making a diagnosis': it is impossible to deal with as many states ('diseases') as there are symptom vectors. One must lump clinical states at the expense of some heterogeneity of potential response to treatment within lumps ('diseases').

The axiomatic basis of the abstract utility concept, to which Fishburn [11] provides a good introduction, has been explained to medical audiences by Lindley [26]. To be useful, this concept has to be 'medicized' like the other components of abstract decision theory. There is no agreement whether and how this can be done. The many difficulties that we have only briefly alluded to (Introduction and Section 6.), are discussed in the several papers on utility measurement cited below; see also Card [7]. Pragmatic approaches to medical utility can be found in, e.g., Alperovitch et al. [2] and the book by Bunker et al. mentioned above.

Whose interests should be quantified in terms of utility? So far, we have assumed, unrealistically, that only the patient's interests matter and that the (individual or typical) patient's perceived loss can somehow be written down in numerical terms so that it can be combined with probabilities based on medical knowledge, using the formal machinery. But the interests of the patient's family, the doctor and other health personnel, future patients, and society at large are all relevant, and often conflicting (see Bunker et al. [6], Chapter 1). The paper of Habbema and van der Maas [14] examines the impact of this conflict. It is concluded that decision theory of today is not entirely appropriate for decision tasks that involve conflicting interests.

But assuming that it is mainly the patient's preferences that matter, the question arises whether the doctor can interrogate the patient without disclosing something of which the patient would prefer to be ignorant (the negative utility of eliciting utilities). A similar remark was made by Savage [38]. In addition, the lay ulcer patient has no way of knowing precisely how unpleasant it would be to get the dumping syndrome; he may in fact never have heard of this after-effect of gastric surgery. He is then likely to follow the leads...
given by his doctor; so in the end it is still the doctor who decides, not the patient. It is somewhat easier to interrogate a patient so as to establish the relative utilities of 1, 2, 3, ... years of survival, as McNeill et al. did in an interesting pilot study of lung cancer patients [28, 29]. However, every patient is likely to attach less importance to the distant than to the near future. Yet to an outsider it may well appear to be equally important to be able to live a normal life next year and ten years from now. Whose then are the utilities after all? Even if the complications are disregarded, preferences over time give rise to difficult theoretical questions to which Meyer (Ch. 9 in Keeney and Raiffa [22]) has worked out an answer. However, he does not succeed in formulating a utility structure that takes into account a patient's worry about his future (nights that are sleepless for fear of the known and the unknown).

Thus, it is not surprising that it is sometimes said that it makes no sense or is impossible to measure medical utilities (and to bring them on a common scale). For those who believe the opposite a number of methods have been suggested in [15, 29, 42], and several of the tutorial papers already mentioned. Non-medical key references include Keeney and Raiffa [22] and a special issue [19] that shows the state of the art. Noteworthy is the work by Card et al. [10], which is likely to attach less importance to the distant than to the near future. Yet to an outsider it may well appear to be equally important to be able to live a normal life next year and ten years from now. Whose then are the utilities after all? Even if the complications are disregarded, preferences over time give rise to difficult theoretical questions to which Meyer (Ch. 9 in Keeney and Raiffa [22]) has worked out an answer. However, he does not succeed in formulating a utility structure that takes into account a patient's worry about his future (nights that are sleepless for fear of the known and the unknown).

First, it is worth noting that, when the regret is set at zero for cases which are correctly classified (by forced classification), otherwise at one loss unit, average actual regret, $Q_{64}$, reduces to the proportion of misdiagnosed cases (error rate); cf. $Q_{6}$. Similarly, $Q_{64}$, which was introduced as a modified non-error rate that pays special attention to rare diseases, is equivalent to $Q_{64}$ with misclassification regrets inversely proportional to the incidence of the actual disease. In Part III, Section 4.2., we argued in favour of using so-called SPSRs (strict proper scoring rules) of which the SPSRs (strictly proper scoring rules) form an especially attractive subset. In brief, a proper scoring rule does not allow a poor diagnosician to beat a good one in the long run. The actual loss/regret statistics ($Q_{6}$ and $Q_{64}$) are SPSRs. So are the non-error rate statistics ($Q_{6}$ and $Q_{64}$) as noted in Part III because they can be viewed as special cases of $Q_{64}$. The variable-regret statistic $Q_{64}$ is in fact an SPSR, as are all scoring rules that are constructed in analogous manner from a suitable regret matrix distribution; for formal treatment, see [31], [35]. Scoring rules for agreement between a predicted distribution on the real number line and an observed numerical outcome (Fig. 3b, Section 5.) can also be classified as improper, proper, or strictly proper; the mathematics is complex. Our misprognostication scores $Q_{6}$ and $Q_{64}$ are SPSRs, but SPSRs also exist (see [27]). That $Q_{64}$ is proper means that an imaginary diagnosician who knows the distribution underlying the outcome cannot, apart from sampling fluctuation, be beaten by any non-perfect diagnosician. That $Q_{64}$ is not strictly proper means that there are non-perfect diagnosciticians that, although they cannot win, are sufficiently close to force a tie (again ± sampling fluctuation). Specifically, if the distributions you predict invariably have the same mean and variance as the perfect prediction, your expected $Q_{64}$ score will be equal to that of the perfect diagnonsitcian.

It should be remembered that the assigned diagnostic probabilities $P_{j}$ can be viewed as estimates of $P_{ij}$ that is, the true large-sample proportion of patients with symptom vector $x_{i}$ that will prove to suffer from $D_{j}$ (Part III, eq. (2)). Were the quantities $P_{ij}$ known, it would be possible to identify, for each patient $H_{j}$, the truly best treatment, say, $T_{ij}$, the associated expected loss being $\Sigma P_{ij}L_{ij}$. The corresponding decision figure for the treatment preferred by the system is $\Sigma P_{ij}L_{ij}$. When averaged over the evaluation sample, the difference between these two figures would tell us exactly how much is lost per patient due to imperfect diagnostics. It may be helpful to note that these expressions contain nothing that is not already in eqs. (1—3); only, $P_{ij}$'s have replaced $P_{ij}$'s anywhere except that it is still the $P_{ij}$'s that determine $r(i)$.

But $P_{ij}$'s are not known. To repeat a point made in Part III (Section 4.1.), the challenge that we try to meet is that of finding ways of rewarding a system whose $P_{ij}$'s are close to the $P_{ij}$'s so that patient loss is close to the minimum possible and, notably, of doing so in settings where $P_{ij}$'s remain unknown and hence cannot be brought into the calculations in person. One may view the actual disease of $H_{j}$ as a proxy of the unknown vector $(P_{1j}, ..., P_{nj})$, the actual loss being $\theta = \Sigma P_{ij}L_{ij}$, an estimate of $\Sigma P_{ij}L_{ij}$. In fact, this is the only sensible estimate available, and it is unbiased because $d(i)$ equals $j$ with probability $P_{ij}$, by the very definition of $P_{ij}$. It is thus indeed rational to use $Q_{64}$ as a measure of «loss-avoiding therapeutic behaviour». The same applies to the regret measure $Q_{64}$ because $Q_{64} - Q_{64} = Q_{64}$, which does not depend on the system but only on patient characteristics.

### APPENDIX A

**A Graphical Analysis of Regret and Choice of Treatment**

The implications of the loss and regret structure of Examples A—C and Section 4. are probably best understood by a study of Fig. 6. It shows an equilateral triangle in which each point represents a possible probability diagnosis $P_{j}$; $(P_{1j}, P_{3j}, P_{5j})$. This type of diagram was extensively discussed in Parts I—II. Using eqs. (7—9) we divided the triangle into four regions, one for each $T_{j}$. For example, the region on the right marked $T_{3}$ comprises all those probability diagnoses that suggest treatment $T_{3}$. Patient $H_{j}$ of Examples B & C falls here ($P_{17} = 4$).

There is indeed a region for each treatment, even for $T_{3}$, which would never be chosen if one knew what was wrong with the patient (no zeros in the third row of the regret matrix). It is left to the reader to verify this. For example, $T_{1}$ is preferred when $P_{1j} = (0.65, 0.30, 0.05)$, i.e. the point marked by $+$ in the diagram.
Fig. 6: Choice of treatment in the clinical situation described by Table 1. The probability triangle has been divided into four regions, one for each treatment. For instance, the probability diagnosis $P_i = (0.65, 0.30, 0.05)$, i.e. the point marked by $T_4$, will suggest that $T_3$ is the treatment of choice, that is to say, the one which entails the least expected regret (eq. (7a)).

The vertices of the triangle represent diagnoses that are absolutely certain in the sense that one disease is assigned 100% probability. As the system effectively states that there is no risk of mistreatment, the minimum expected regret $e_i$ is zero. In this particular application, $e_i$ is also zero along the entire $D_2-D_3$ side of the triangle which is seen to be part of the $T_2$ region, because immediate operation is a safe decision when either $D_1$ or $D_2$ must be present: The fact that $k_{32} = k_{33} = 0$ ensures that, whenever $D_3$ can be ruled out, the doctor is certain not to regret choosing $T_2$. The expected regret is in fact less than one unit everywhere outside the contour marked $e_i = 1.0$. The expected regret contours, a few of which are shown, are straight lines within each $T_r$ region because $e_{ri}$ is a linear function of the three assigned probabilities, cf. (7a). Moreover, for $P_i$ to fall on the $T_1-T_2$ boundary, say, the assigned probabilities must satisfy the linear condition $e_{r1} = e_{r2}$. Hence the boundary becomes a straight line segment (along which the expected regret contours will kink). The treatment regions, therefore, become polygons (eq. (9) in fact implies that they are convex). They meet at three-region corner points where three $T_r$'s are equally good and better than the fourth.

One of the corners is $S = (247/396, 95/396, 54/396)$. As visualized by the contour lines, the largest expected regret is attained there. The reader may verify that $e_i = e_{r1} = e_{r2} = e_{r4} = 2470/396 = 6.24$ which is less than $e_{r3} = 2870/396$. This is, therefore, the least lucky point for a patient to fall into in the sense that no matter what the doctor decides to do the expected regret is at least 6.24 units. As noted in Example C, the largest expected regret encountered in the Acute Abdominal Pain data is that of $H_{17}$ ($e_i = 5.03$). This patient also happened to have the largest actual regret (13 units) because somewhat unexpectedly he turned out to have $D_3$, but that is a different matter.

It is obvious from the $D_3$ column of Table 1.1b why the $T_2-T_3$ boundary passes through vertex $T_3$.

By definition, the indifferent system (Section 3.) will produce predictions located at the centre of the triangle. In agreement with Table 3 this point is seen to belong to the $T_2$ region, the associated minimum expected regret being $e_i = 3.33$ (cf. contour lines). This figure reappears in Table 6 as $Q_{65}$ (average minimum expected regret). Analogous remarks apply to the incidence system, the predictions of which fall at the point marked $T$, being again constant across patients and now equal to 3.40.

From a purely probabilistic point of view, the centre is the point of maximum uncertainty because the diseases are judged equiprobable. From a therapeutic point of view, it is point $S$ that involves a maximum of diagnostic uncertainty because the loss structure of Table 1 implies that this is the point of maximum expected regret, as already discussed.

It is easy, once the diagram of Fig. 6 has been drawn, to superimpose on it the data of the individual patients in the manner indicated in Fig. 7 which we find self-explanatory and instructive (cf. Fig. 1).

Fig. 7: Acute Abdominal Pain data. Regrets superimposed on the probability triangle. (All but eleven regrets are zero.) The diagram, when compared with Fig. 6, shows not only the severity of the 11 treatment-selection errors displayed in Fig. 1, but explains also how they arose as a logical consequence of probability misallocation. The maverick nature of case No. 17 leaps to the eye. The main part of the regret load is seen to be unavoidable unless a substantially better separation of $D_1$ and $D_3$ is achieved.

We are now studying the case of $k = 3$ diseases. With only two diseases the analogues of Figs. 6 and 7 are still more convenient. Let us give a few hints: $e_i$ becomes a tent-shaped function like the deterministic curve ($\beta = \infty$) in Fig. 4, lower part, and individual regrets follow the pattern of the corresponding step curves in the upper part of that figure.
APPENDIX B
Tests of Reliability
A test of the null hypothesis of perfect reliability answers the question whether or not a discrepancy between an actual and an expected score is significant. Similarly, the deviation of one or zero accordingly as the latter is one or zero is tested by dividing

\[
\text{dev}^2(Q_{34}) = Q_{34} - \frac{Q_{35}}{Q_{35} + Q_{35}} = +0.14
\]

by the associated standard deviation,

\[
\frac{1}{N} \sum_{i=1}^{N} (L_{0i} - E_i)^2 = 1.05.
\]

The resulting approximately standard-normal statistic becomes +0.13 and is obviously not significant. The corresponding test of \(\text{dev}^2(Q_{35})\) yields a value of +1.14 (non-significant difference). Similarly, the deviation 

\[
\frac{1}{N} \sum_{i=1}^{N} (Q_{35} - C_{i})^2 = 1.05
\]

is found to be only -0.86 times the associated standard deviation.

It should be borne in mind that with only 50 patients in the sample there is little power to detect moderate departures from perfect reliability.

APPENDIX C
The Ranked Quadratic Scoring Rule
In the situation described at the end of Section 5.2, evaluation of predictions must be based solely on making a vertical cut through the step curves (predicted and actual) of Fig. 3a in each of the \(k-1\) intervals between neighbouring \(L_s\)'s. The ordinate of the former curve represents the predicted probability of the event \(\leq L_0\) is lower than this cutpoint, while the latter is one or zero according as \(L_0\) is actually lower or higher than the cutpoint. Once this is done, any two-outcome scoring rule can be applied interval-wise to the alternative outcomes \(\leq L_0\) is higher and \(L_0\) is lower, the individual interval scores being added to form a final score. This will then be a measure of prediction-outcome agreement in which the ordering of the \(k\) alternatives has been taken duly into account. In particular, if quadratic scoring \((Q_{34})\) is employed, one obtains the ranked quadratic scoring rule

\[
\text{RQSR} = Q_{34} = \frac{1}{N} \sum_{i=1}^{N} (G_{ij} - A_{ij})^2, \quad (49)
\]

where \(G_{ij}\) is the ordinate of the predicted curve in the \(j\)th interval from the left and \(A_{ij}\) is the corresponding ordinate for the actual curve, equal to one if \(L_{0j}\) is any of the \(j\) least serious \(L_s\)'s, otherwise zero. The best value for \(Q_{34}\) is zero, and the worst possible score is \((k-1)\). With the Abdominal Pain data the scores are: for the actual system 0.241; for the indifferent system 0.502; for the incidence system 0.457; theoretical maximum 2.0. The theory of this scoring rule is explained in [40].

References

\[
\]


Addresses of the authors: J. D. F. Habbema, Dept. of Public Health and Social Medicine, Erasmus University, P.O. Box 1738, Rotterdam, The Netherlands and Dept. of Medical Statistics, University of Leiden, Wassenaarseweg 80, Leiden; J. Hilden, Univ. Inst. of Human Genetics, Tagensvej 14, DK—2200 Copenhagen, Denmark.