Reassessing Accuracy Rates for Median Decision Procedures

Andrea Capotorti and **Frank Lad**
University of Perugia University of Canterbury

Key Words: Asbestosis, conditional exchangeability, second opinions, medical diagnosis, specificity, sensitivity, predictive values, probability bounds, fundamental theorem of prevision, linear programming, quadratic programming

Abstract

We re-examine the procedure of median decision making in the context of radiological determination of asbestosis by three B-readers. Our assessment addresses the specificity, sensitivity and predictive values from this procedure compared to merely an individual radiologist's diagnosis. Conditional exchangeability of the radiologists' classifications is recognised as more appropriate than independence which is often presumed. The framework of de Finetti's fundamental theorem of probability makes the analysis tractable when it is formulated in terms of a linear programming problem, yielding coherent bounds on probabilities of interest even when a complete distribution is not specified. Further natural assertions motivate a partial ordering of conditional probabilities. In this context the computation of bounds develops into a quadratic programming problem. Using sensible assertions, the median decision procedure is found to be relatively weaker than has been thought based on the presumption of independence of radiologists' assessments. However that presumption is also shown to overstate the predictive qualities of individual diagnoses. We re-evaluate substantive claims about the use of median X-ray decisions as an indicator of cancer.

1 Introduction

The willingness of statisticians to presume the stochastic independence of quantities in order to achieve an analytic result in an applied problem, even when this presumption is inappropriate, can result in misrepresentation of the information contained in recorded data when it is used to make inference. A striking example of this situation arises when the data are probability assertions or other judgments by experts regarding the condition of a patient. In this article we consider specifically the classifications of a patient by three radiologists on the basis of an X-ray that is indicative but not definitive about the state of the patient's lung tissue. Dependencies among the assessments of several experts are built into their judgments because they have each qualified in some public way as a registered expert, typically by passing a standard examination. Since each of them has received similar training and has passed the same examination in order to achieve certification, anyone who uses their judgments would typically vary his/her probability for a second expert's classification depending on the stated judgment of the first expert's reading. To statisticians who are familiar with the concept, it would be natural to consider that an appropriate structure of exchangeability should be built into the analysis of experts' judgments when using them to make final diagnosis and intervention decisions.

An important article on this problem by Tweedie and Mengersen (1999), hereafter T-M, makes use of the questionable independence presumption we have just mentioned. Their aim was to quantify the accuracy of a decision procedure in which three medical practitioners together determine the diagnosis of a patient's lung condition when at least two of the three of them judge that the evidence from the X-ray exceeds a defined minimal critical level, or not. This decision procedure, termed *median diagnosis*, is used to establish fibrosis of the lung associated with friable asbestos exposure (asbestosis) according to an internationally specified standard. In order to achieve an estimable result, T-M were willing to presume as independent the successes of the three experts in correctly assessing patients who have the condition, and similarly as independent when assessing patients who do not have the condition. They made this assumption because the radiologists make their personal determinations upon viewing the X-ray without knowledge of the other two radiologists' judgments. The presumption is much more widespread than merely this article, however. Walter and Irwig (1988) for example, presented an extensive review of related research which universally relied on it. Although they considered this unfortunate, they found few obvious practical alternatives leading to computable results. The article of T-M itself querries how to insert into the problem the condition that some cases are easier to diagnose than others. This would mean that the probability of a third assessor diagnosing the condition increases on the condition that the other two have done so, even when the third assessor does not know the conclusion of the other two's judgments.

In this article we shall propose that a form of conditional partial exchangeability is appropriate to the analysis of the experts' judgments, and provide a tractable resolution to the analysis. Related work in Bayesian statistics has addressed widespread problems of diagnosis on the basis of correlated evidence but without the availability of a gold-standard test. From the proposals of Geisser and Johnson (1992) through to Black and Craig (2002) analysis has been based on mixture parametric distributions over various forms of conditional dependencies. As will be seen in Section 2, the practical context of the fibrosis problem precludes honest assessment of complete prior distributions. Thus, our analysis differs from these developments in that it does not require a completely specified distribution. Relying on limited supported assertions, it computes bounds on probabilities of interest much in the spirit of Walley's (1991) proposed imprecise probabilities. Computations based on the linear programming structure of de Finetti's *fundamental theorem of probability* (Bruno and Gilio, 1980; Lad, 1996, 2.10, 3.3) yield interesting numerical results to compare with those of T-M, and motivate rather different conclusions regarding median decision procedures. We provide a complete comparison of aspects of median decisions that have not been assessed to date.

In Section 2 we discuss some practical matters relevant to the radiological diagnosis of asbestosis and to the problem studied by T-M. Section 3 presents details of the logical relations among the central events involved. In Section 4 we review the goals of the inferential analysis and the technical details of the work presented by T-M. We then describe how to formalise the alternative conditional exchangeabilities that would be natural to presume among the experts' assessments. We also outline some inequalities on conditional probabilities that are appropriate to the problem. In Section 5 we present computations of bounds for inferential statements regarding the specificity, sensitivity and predictive values of median decision procedures using these presumptions, and we discuss several comparisons of interest. Section 6 describes the computational methods we used in computing the coherent intervals under the array of constraints we motivate, and it describes some recent advances in computational applications of de Finetti's fundamental theorem of prevision. We conclude with discussion of substantive aspects of asbestosis issues in Section 7.

2 X-ray diagnosis of asbestosis

The assessment of radiological evidence regarding asbestosis of the lung is extremely complicated for political, economic and judicial reasons, as well as for medical statistical ones. The easiest way for an interested reader to learn about a wide array of issues involved would be to read the presentation of an experienced pulmonary physician (Martin, 2002) who has posted a very readable commentary on the web. We do not intend to address many of these reasons in this article. Rather we are focusing expressly on the probabilistic properties of a particular diagnostic strategy that has been termed a *median decision procedure*. This group-diagnostic procedure has a wider range of applicability than the specific problem of asbestosis. Thus, the technical properties of the procedure are worthy of study in themselves. However, in order to explain the motivation for the use of this procedure and the specific context of its application to the study of fibrosis of the lung, we present here some of the relevant background.

Asbestos is a mineral composed by metalic structures of magnesium, silicon and iron in various forms. Individual fibres cannot be seen directly through a light microscope. However, when small particles of asbestos dust are inhaled into lung tissue, after a long process of at least 20 years they come to be coated with ferrous material. The existence of these covered bodies (which constitutes the condition of asbestosis) can often be inferred from an X-ray that exhibits a marked shadowing on the film. A detected shadow is referred to as a *small opacity*. The detection of such ferruginous bodies by a radiologist is quite difficult, even by very experienced readers. In order to qualify as an assessor of an X-ray for this corrupted lung condition of asbestosis, a radiologist is required to pass a specific accrediting examination. For historical reasons, radiologists so accredited are called *B-readers*.

An international standard has been developed for a B-reader to classify the status of a pulmonary X-ray based on an assessment of the number of small opacities per unit surface area shown on a film. On the basis of a procedure specified operationally by the International Labour Office in 1980, the radiologist classifies the X-ray into one of 12 possibilities, ordered from the least damaged lung to the most damaged. The classification categories make up a 12 point scale, designated 0/-, 0/0, 0/1, 1/0, 1/1, 1/2, 2/1, 2/2, 2/3, 3/2, 3/3, 3/+. The first number in each separated pair represents the judgment to classify the film into one of four major categories (0, 1, 2 or 3) by comparing it to standard radiograph pictures, ranging from "no small opacities" to a very high density. Upon viewing an X-ray, a reader will typically be somewhat uncertain as to which category is pertinent, but must summarise the reading by one of these categories. Recognising this difficulty, the reading procedure requires that a second classification must be specified by the same reader. The second number after the slash is a refinement which specifies whether or not the next lower or higher major category was seriously considered in the first classification judgment.

How extreme should the evidence from an X-ray be to diagnose the condition of asbestosis of the lung radiologically? The criterion specified by the American Thoracic Society in 1986 was that the X-ray should fall in a category at least as extreme as 1/1. However a group of experts in 1997 resolved that a weaker standard of evidence, known as the Helsinki criteria, should recognise the classification of 1/0 as also supporting evidence of an early stage of asbestosis.

Arguments over the severity required have been complicated by still another matter. Even after a considered viewing of an X-ray, a B-reader may well be uncertain as to the status of a patient's lung, and a different reader with similar level of accredited skill may judge it to be in another category based on the same X-ray. Thus, a procedure has been established requiring that an official diagnosis of asbestosis is recognised only if two out of three different B-readers assess the X-ray with a classification of 1/0 or worse. This is the procedure that formally defines a *median diagnosis.*

The condition of asbestosis manifests itself clinically by shortness of breath and pain upon deep breathing, among other things. Physical activity becomes restricted as the condition worsens. The only unequivocal procedure for identifying the presence of ferruginous bodies in the lung is a histological examination of a sample of lung tissue. However, since there is no treatment that can alleviate the condition, such an invasive examination is virtually never made while a patient is alive. Most evidence on the relation between radiological analysis of X-ray films and histological results is collected from autopsies conducted when the cause of death is designated as lung cancer. This feature of the situation will be recognised in the next Section as putting limitations on data collection that would be relevant to a complete statistical analysis of evidence designed to improve our understanding.

3 Event structure for median diagnosis

We shall denote by F the event that a specific exposed patient has the condition of fibrosis of the lung. Each of three radiologists examine a pulmonary X-ray of the patient, and assess it in one of the categories specified by the international standard. Let D_i denote the event that radiologist i assesses the film with a category at least as extreme as 1/0. There is no logical restriction in these definitions that prevent any of the sixteen possible outcomes of the vector of the four events we have just defined: F, D_1, D_2 and D_3. Thus, we say these events are *logically independent.* This feature of the events is quite distinct from the question of their probabilistic dependence, which is another matter completely.

We shall begin our analysis by exhibiting the realm matrix $\mathbf{R}(\cdot)$, of possible vectors of values for several events that interest us in this problem, and then discuss its construction. Our notation follows de Finetti (1967) in identifying events with their indicator variables, that is as numbers rather than as sets. Thus for exampe, the event D_1 equals either 1 if the first reader's diagnosis is positive, or 0 if it is negative. Using this convention, events can be summed or multiplied. The negation of an event such as D_i is defined numerically by $\widetilde{D_i} \equiv 1 - D_i$. We also use the convention that a parenthetical expression containing an arithmetical statement that could be true or false denotes the event that equals 1 if the statement is true, and equals 0 if it is false. For example, $D^* \equiv (\sum D_{i=1}^3 \geq 2)$ defines the event that at least two of the radiologists' diagnoses are positive. Here now is the promised realm matrix.

$$\mathbf{R}\begin{pmatrix} F \\ D_1 \\ D_2 \\ D_3 \\ \text{- - - - - - - - -} \\ \sum_{i=1}^{3} D_i \\ D^* \equiv (\sum D_{i=1}^{3} \geq 2) \\ S^* \\ \text{- - - - - - - - -} \\ \sum_{i=1}^{3} D^*(1-D_i) \\ S^* D^* \\ \text{- - - - - - - - -} \\ D_3\widetilde{D_2}\widetilde{D_1}\widetilde{F} \\ \widetilde{D_1}\widetilde{F} \\ D_3\widetilde{D_1}\widetilde{F} \\ \widetilde{D_2}\widetilde{D_1}\widetilde{F} \\ D_1\widetilde{D_2}F \\ D_1\widetilde{D_2}\widetilde{F} \end{pmatrix} = \begin{pmatrix} 0&1&0&0&0&1&1&1&0&0&0&1&1&1&0&1 \\ 0&0&1&0&0&1&0&0&0&1&1&0&1&1&1&1 \\ 0&0&0&1&0&0&1&0&1&0&1&1&0&1&1&1 \\ 0&0&0&0&1&0&0&1&1&1&0&1&1&0&1&1 \\ -&-&-&-&-&-&-&-&-&-&-&-&-&-&-&- \\ 0&0&1&1&1&1&1&1&2&2&2&2&2&2&3&3 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&0&0 \\ -&-&-&-&-&-&-&-&-&-&-&-&-&-&-&- \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1&1&1&0&0 \\ -&-&-&-&-&-&-&-&-&-&-&-&-&-&-&- \\ 0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&1&1&0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0 \\ 1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0&1&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0 \end{pmatrix} \qquad (1)$$

The sixteen columns of the first four rows of this realm matrix identify the exhaustive list of possible situations of the four events we have defined. Of course the order in which we list the columns of their possibilities is irrelevant to the logic of the analysis or its computational development. However, we present them in an order that is hopefully conducive to their understanding and our subsequent discussion.

Each element of the first row of the realm matrix identifies by 0 or by 1 whether the considered lung is not or is in the condition of fibrosis. The next three row elements in each column specify possible results of the three B-readers' assessments of an X-ray film. In the first column, for example, the zero in the first row specifies that the lung is not in the condition of fibrosis, while the next three rows identify the result that all three of the radiologists assess the X-ray in a category less extreme than 1/0. This same result of their diagnoses appears in the second column too. However, the fact that the first row component of the second column equals 1 identifies that the patient so diagnosed as negative by all three radiologists actually *does* have the condition of fibrosis of the lung. Of course this is a possibility.

The next three columns of the matrix begin with 0 in their first row, and are followed by one 1 and two zeros in their three possible permutations. In each case, one of the radiologists gives a diagnosis of 1 and the other two a 0, but the radiologist who makes the dissenting diagnosis is different in each case. So these columns designate the possibilities that the patient does not have fibrosis, and the median decision procedure correctly arrives at this diagnosis by a majority but not unanimous decision. Similarly, the following three columns (6-8) exhibit in their first row by the 1 that the patient *does* have fibrosis, while the next three rows again exhibit the three possible permutations of one diagnosis of 1 and two of 0. In each case, an incorrect median diagnosis is reached by a split decision. The realm matrix of possibilities continues with the columns ordered in this way. Columns 9-11 show no fibrosis, since F = 0, but the median diagnosis is in favor of fibrosis by a majority but not unanimous decision, and so on.

The remaining rows of the matrix identify the values of various functions of these sixteen constituent possibilities. They are partitioned into blocks merely for expository purposes. Row five, in the second partitioned block, identifies the sum of individual diagnoses in favor of the condition. Row six identifies the event that the median diagnosis is in favor of fibrosis or not, denoted by $D^* \equiv (\sum_{i=1}^{3} D_i \geq 2)$. Row seven identifies the event that a positive median diagnosis is achieved only by a split vote, not a unanimous decision. Algebraically, it is defined by $S^* \equiv (\sum_{i=1}^{3} D_i = 2)$.

The third partitioned block contains only two quantities. The first is the sum of the events that each of the individual readers disagrees with a positive median decision. These individual events are expressed algebraically by $D^*(1-D_i)$ for each i. It is evident that the sum of these three events is identical to the event S^*, seen in the preceding row. T-M used this result to assess a probability for S^* on the basis of data on the extent of disagreement of individual assessors with the positive median decisions. We shall specify their assessed probability value shortly. The second quantity displayed in the final block, S^*D^* is again derived merely from the component multiplication of the rows of the realm for S^* and D^*. It is equivalent to the row for the event S^*. We display it here only because it will be used in specifying another probability as assessed by T-M in what follows.

The final partitioned block of six events will be relevant to discussion at the end of Section 4. Hopefully you can notice how the 1's in each row identify those columns for which the $(F, D_1, D_2, D_3)^T$ configurations imply the occurrence of the event identifying that row.

4 Accuracy of median decisions

A complete probabilistic assessment of lung condition and diagnostic possibilities for an individual patient would be specified by 16 probabilities, one for each column of the realm matrix we have specified in (1). In the remainder of this article we shall refer to these probabilities by the notation of a column vector, $\mathbf{q}_{16} \equiv (q_1, q_2, ..., q_{16})^T$, which would include a probability for each of the possibilities. Unfortunately for our understanding of the median decision making procedure in this instance, it is very difficult to gain enough information about some of these possibilities to assert precisely the required probabilities. Particularly for the constituent events defined by columns 1-8, information is very seldom procured about the lung condition of patients for which the median diagnosis is negative – that is, when $D^* = 0$. Nonetheless, there is evidence available about some aspects of this problem, and there is a widely agreed upon structure to professional opinion about other aspects. We shall capitalise on these assertions to derive bounds on the probabilities relevant to characterising the median decision procedure that we desire. In this Section we shall provide details of informational inputs for the application of Bruno de Finetti's fundamental theorem of probability that allow us to derive these bounds. We shall present a brief statement of the theorem in its computational form in Section 5 before we state the results.

4.1 What probabilities do we require?

In any diagnostic problem assessed by an individual physician, there are four conditional probabilities that are widely considered to characterise the accuracy of the physician's expected diagnostic performance. Standard terminology coins them as the *sensitivy, specificity, positive predictive value,* and *negative predictive value* of a diagnosis. Specified in terms of an individual radiologist in this problem, these are the probabilities $P(D_i|F), P(\widetilde{D_i}|\widetilde{F}), P(F|D_i), \text{and} P(\widetilde{F}|\widetilde{D_i})$. Ideally, all four of these probabilities would be large, as close as possible to 1 in each case.

The goal of our analysis is to identify the difference between these characteristic diagnosis probabilities for individual B-readers and the corresponding diagnosis probabilities for the median decision procedure. These are denoted by $P(D^*|F), P(\widetilde{D^*}|\widetilde{F}), P(F|D^*)$, and $P(\widetilde{F}|\widetilde{D^*})$. Since median diagnosis is based on stricter standards, we expect each of these probabilities to exceed those for an individual reader's diagnosis. Of interest is the size of the gain achieved by requiring the median procedure in any instance.

4.2 Exchangeability restrictions

There are two aspects of radiological assessments of lung X-rays that motivate exchangeable judgments. Firstly, there is a symmetry to the conditions of patients who require an X-ray of their lungs for purposes of diagnosing asbestosis. Patients collected for published studies are usually chosen on the basis of a standard condition that they have a working history of exposure to friable asbestos beginning more than twenty years before the exam, and they exhibit some clinical evidence of correlated symptoms such as shortness of breath on a lung capacity test. The judgment of exchangeability among the patients' lung conditions amounts to an assertion of equal probabilities to every sequence of n patients' conditions that exhibits the same number of positive fibrosis cases, no matter what the order in which the positives and the negatives occur. There would be little disagreement with this assertion among trained statisticians. The judgment of exchangeability is what allows us to learn about the prevalence of asbestosis diagnoses among some patients in a well-defined group on the basis of the observed results of X-ray diagnoses for others in the group. This symmetry of attitudes toward patients' conditions was suggested by T-M. The other aspect motivating exchangeability is the symmetry of the judgments by accredited radiologists who examine the same X-ray. We expand its treatment beyond the T-M proposal in the next two subsections.

4.2.1 Conditional independence ?

T-M's treatment of exchangeability of diagnoses among radiologist B-readers follows common practice, but it is open to improvement, so we shall first describe their approach. They suppose that the diagnoses by the three examining radiologists are conditionally independent both given the actual condition of fibrosis, F, and also given its negation, $\widetilde{F}$. Technically, this does amount to an assertion of exchangeability regarding the events D_1, D_2 and D_3, because a mixture of conditionally independent distributions is an exchangeable distribution. However, this condition is too confining to represent the full extent of the symmetries most analysts would find appropriate to this problem.

This presumption implies that the 16 elements of $\mathbf{q}_{16}$ would be determined by two probabilities which they denote by $p \equiv P(D_i|F)$, and $p_f \equiv P(D_i|\widetilde{F})$ along with a third probability, $P(F)$. The probabilities p and p_f are presumed to be equal across assessors, and the corresponding events of their positive diagnoses, D_1, D_2 and D_3, are specified as conditionally independent. With such a presumption, the probabilities for columns 1, 3-5, 9-11 and 15, and columns 2, 6-8, 12-14 and 16 of the matrix in (1) would be generated by two distinct conditionally Binomial distributions.

For reasons that we have discussed, the direct assessment of the probabilities p, p_f and $P(F)$ is difficult. However, the conditionally Binomial mixture probabilities put so much structure onto the problem that two different and accessible probabilities can aid the identification of the details. T-M use this structure to derive a equation relating the probabilities p and p_f to one another. To prevent distraction, we discuss the details of the equation in a brief Appendix for those interested. Suffice it to say here that their development relies critically on the independence of the radiologists' diagnoses given F and given $\widetilde{F}$. Its numerical solution uses the specification of values for $P(D^*)$ and $P(S^*|D^*)$, or equivalently $P(S^*D^*)$ to provide possible values of the pair (p, p_f). T-M supply the required assertion values as $P(D^*) = .12$ and $P(S^*|D^*) = .42$, or equivalently, $P(S^*D^*) = .0504$, on the basis of statistical evidence which they reference. Finally, they consider the implications of a range of values for p and the companion p_f it would "imply", including $p = .82, .90$ and $.96$. The extremes of this triple were suggested by histological studies which they also reference, and the interior value of .9 was promoted by their discussion of the computational results.

Histological examinations are typically not conducted for a live presenting patient, nor are autopsies conducted with this focus when the median decision procedure is negative. Thus, little statistical evidence is available about the constituent events defined by columns 1-8, while evidence relevant to columns 9-16 is limited. Thus, T-M prise their derived equation relating p to p_f since it allows them to compute corresponding values "implied" for p_f on the basis of an array of possible estimates of p. They then use these pairs to compute positive predictive values and negative predictive values of the median decision procedure which they publish and discuss. In a discussion, T-M lament that their assumptions will not allow them to account for the fact that some X-rays are easy to assess while others may be difficult. This would mean that the Binomial probability evaluations are inappropriate. Let us now look into how the computable implications of this recognition can be addressed.

4.2.2 Conditional exchangeability

It is generally recognised that the susceptabilities of an X-ray to exhibit true and false clues when the patient has fibrosis of the lung and when the patient has not this condition are quite different matters. Thus, consideration of probabilities such as $P(D_1D_2\widetilde{D}_3|F)$ and $P(D_1D_2\widetilde{D}_3|\widetilde{F})$ are different endeavors as well. However, since the radiologists are subject to standard training and an accreditation examination, it is *at this conditional point* that the judgment of exchangeability should be made. B-Readers' diagnoses should be regarded exchangeably given F and also exchangeably given $\widetilde{F}$. What this means technically is firstly that the components of $\mathbf{q}_{16}$ corresponding to columns 3-5 should be equal since these columns are observationally identical except for the fact that the one incorrect positive diagnosis among the three radiologists' judgments is made by a different reader in each column. Similarly, columns 6-8 involve only a permutation of the order of the single correct diagnosis among the three readers. Thus the probabilities $q_6, q_7,$ and q_8 should be equal. The same symmetry arguments would require the equating of probabilities associated with columns 9-11 and columns 12-14. This recognition of conditional exchangeability of the radiologists' judgments given F and given $\widetilde{F}$ amounts to a partial exchangeability of their judgments irrespective of the condition of the patient's lung.

We are left to the problem of identifying how this condition can be input into a linear programming problem to implement de Finetti's fundamental theorem of probability, determining the cohering probability bounds for events that concern us.

4.3 Linear programming inputs

To begin, notice that the asserting conditional exchangeability would put the following constraints on a cohering distribution over the 16 columns of the realm matrix, denoted by the mass-function vector $\mathbf{q}_{16}$:

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \text{------} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} & \text{--} \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \mathbf{q}_{16} \tag{2}$$

These first eight row equations ensure that appropriate components of $\mathbf{q}_{16}$ are equal. (For example, row 1 ensures that $q_3 = q_4$.) The partitioned row 9 ensures that the components of $\mathbf{q}_{16}$ sum to 1. Of course there are many vectors $\mathbf{q}_{16}$ that would satisfy these nine linear constraints.

These restrictions will be supplemented by numerical constraints on $P(D^*)$ and $P(S^*|D^*)$ motivated by the literature search of T-M: $P(D^*) = .12$ and $P(S^*|D^*) = .42$, implying $P(S^*D^*) = .0504$. These two constraints are represented by the further constraint matrix equation

$$\begin{pmatrix} .12 \\ .0504 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix} \mathbf{q}_{16} \tag{3}$$

In all, equations (2) and (3) provide eleven linear constraints on $\mathbf{q}_{16}$.

4.4 Further quadratic constraints

The similarities of the training and qualifications of the B-reader radiologists support further a partial ordering of conditional probabilities on the three separate diagnoses. We shall display them in a sequence of rows of inequalities, motivating them row by row. In concluding this Section we shall notice that these inequalities specify quadratic constraints on the components of $\mathbf{q}_{16}$.

$$P(D_3|\widetilde{D_2}\widetilde{D_1}\widetilde{F}) \;\leq\; P(D_3|\widetilde{D_1}\widetilde{F}) \;\leq\; P(D_3|\widetilde{F}) \;\leq P(D_3|F) \;\leq\; P(D_3|D_1F) \;\leq\; P(D_3|D_2D_1F) \tag{4}$$

Consider this first row of inequalities *from right to left.* To begin, these inequalities express the realisation that in the context of a patient who has fibrosis, the condition of positive diagnosis by a B-reader entices us to a greater expectation of positive diagnosis by the next reader had we not been privy to this condition. As we pass each inequality from right to left we are conditioning on one less positive diagnosis each time. The middle inequality expresses the view that positive X-ray diagnosis of a patient with fibrosis is assessed with higher probability than a positive diagnosis for a patient without fibrosis. This merely expresses the confidence we have that X-ray examination is useful for diagnosis of fibrosis. Continuing the inequalities to the left, we express the contrapositive opinion that a negative diagnosis by a reader of a patient who does not have fibrosis provides evidence for us to expect more strongly that the next diagnosis will be negative as well. Recall, when examining the inequalities in (4), that the previously asserted symmetries in our attitudes toward the three B-readers would imply that we could permute the subscripts in any inequality without affecting the status of the inequality.

The next rows of inequalities shall be presented as a pair:

$$\begin{aligned} P(D_2|\widetilde{D_1}\widetilde{F}) \quad &\leq \quad P(D_2|\widetilde{D_1}F) \quad \leq \quad P(D_2|F) \qquad \text{and} \\ P(D_2|\widetilde{F}) \quad &\leq \quad P(D_2|D_1\widetilde{F}) \quad \leq \quad P(D_2|D_1F) \quad . \end{aligned} \tag{5}$$

Reading from right to left again, the first row expresses that conditioning on an incorrect negative diagnosis diminishes our expectation of a positive diagnosis by the second reader; and our expectation of a positive diagnosis by the second reader is even further diminished when conditioning on a correct negative diagnosis. The companion second row of inequalities merely expresses the contrapositive of these relations.

The reason the inequalities of (5) are separated from those of (4) is that it is difficult to know how to order the relative sizes of some of the conditional probabilities that appear in different lines. Consider the relative sizes of $P(D_2|\widetilde{D_1}F)$ and $P(D_3|\widetilde{F})$, for example. Is the content of the signal of a negative diagnosis on a patient with fibrosis stronger or weaker regarding the next diagnosis than the mere condition of not having fibrosis? Without evidence that is hard to accumulate, it is better to leave this relation unspecified than to make some arbitrary judgment. We shall learn from our computations whether coherence with the other assertions that we can make puts any restrictions on the direction of this inequality. The placement of $P(D_2|D_1\widetilde{F})$ from the second inequality row of (5) relative to $P(D_2|F)$ is similarly problematic.

The next affirmed inequalities also should be viewed as a pair:

$$\begin{array}{lllll} P(D_3|\widetilde{D_2}\widetilde{D_1}F) & \leq & P(D_3|\widetilde{D_1}F) & \leq & P(D_3|\widetilde{D_1}D_2F) \quad \text{and} \\ P(D_3|\widetilde{D_2}D_1\widetilde{F}) & \leq & P(D_3|D_1\widetilde{F}) & \leq & P(D_3|D_2D_1\widetilde{F}) \quad . \end{array} \tag{6}$$

These inequalities are difficult to order within the structure of (5) on account of a problematic relation between the end points of the two inequalities. Compare, for example, the left-side elements of the first rows of (6) and (5): $P(D_3|\widetilde{D_2}\widetilde{D_1}F)$ with $P(D_2|\widetilde{D_1}\widetilde{F})$. Both of these are expected to be smaller than $P(D_3|\widetilde{D_1}F)$ which appears to the right of each. But which of them should be larger? Would the condition of two negative diagnoses of a patient with fibrosis make you expect the third diagnosis to be positive more or less strongly than would the condition of only one negative diagnosis but on a patient without fibrosis? Again this is a difficult question for which coherency with more convincing inequalities will be of interest. On the right side appears a similar conundrum, the relative sizes of $P(D_3|\widetilde{D_1}D_2F)$ and $P(D_2|F)$. Similar difficulties arise when comparing probabilities in the second rows of (5) and (6).

Finally, another array of inequalities can be ordered about the value of .5 :

$$P(D_3|\widetilde{D_1}\widetilde{F}) \;\leq\; P(D_3|\widetilde{D_1}D_2\widetilde{F}) \;\leq .5 \;\leq\; P(D_3|\widetilde{D_2}D_1F) \;\leq\; P(D_3|D_1F) \quad . \tag{7}$$

Comparing the first row of inequalities in (6) with (7) it seems difficult to know just where to place the value of .5 within the ordering of (6) whereas it is natural to order it between $P(D_3|\widetilde{D_1}D_2\widetilde{F})$ and $P(D_3|\widetilde{D_2}D_1F)$ in (7).

Having specified 17 further inequalities via the partial orderings of inequalities (4 - 7), we now should notice that 15 of them amount to quadratic constraints on the components of $\mathbf{q}_{16}$, while two of them are linear. Consider the first inequality of (4), for example, $P(D_3|\widetilde{D_2}\widetilde{D_1}\widetilde{F}) \;\leq\; P(D_3|\widetilde{D_1}\widetilde{F})$. This is an inequality on two conditional probabilities where the conditioning events are different from one another. Multiplying both sides by the product $P(\widetilde{D_2}\widetilde{D_1}\widetilde{F})P(\widetilde{D_1}\widetilde{F})$ then yields the equivalent product inequality

$$P(D_3\widetilde{D_2}\widetilde{D_1}\widetilde{F})P(\widetilde{D_1}\widetilde{F}) \;\leq\; P(D_3\widetilde{D_1}\widetilde{F})P(\widetilde{D_2}\widetilde{D_1}\widetilde{F}) \quad . \tag{8}$$

Each of the multiplicand probabilities on the two sides of (8) are expressible as linear functions of $\mathbf{q}_{16}$. Their representation can now be recognised in the final partitioned block of the realm matrix shown in equation (1). Thus, the product inequality constitutes a quadratic inequality on $\mathbf{q}_{16}$:

$$q_5\ (q_2 + q_7 + q_8 + q_{12}) \quad\leq\quad (q_5 + q_9)\ (q_1 + q_5) \quad . \tag{9}$$

Of the 17 inequalities listed in this Section, 15 of them have this property of being quadratic inequalities on the components of $\mathbf{q}_{16}$. The two inequalities surrounding the number .5 in (7) amount only to linear inequalities on $\mathbf{q}_{16}$, because for example the assertion of $P(D_3|\widetilde{D_1}D_2\widetilde{F}) \;\leq\; .5$ is equivalent to $P(D_3\widetilde{D_1}D_2\widetilde{F}) \;\leq\; .5\ P(\widetilde{D_1}D_2\widetilde{F})$. This inequality is linear in $\mathbf{q}_{16}$.

In the next Section we shall display the numerical results that derive from these inequalities and the conditional exchangeability assertions.

5 Quadratic programming results

Some problems of probability can be resolved to a unique solution on the basis of other probabilities that can be specified. The problem we are examining here is a prime example of a problem for which the information available is insufficient to resolve uniquely the probabilities that interest us. Bruno de Finetti characterised the boundaries on coherent solutions to such problems as early as his famous lectures in Paris (de Finetti, 1937), and he named this result "the fundamental theorem of probability" in his text of 1974 and 1975. An extensive presentation of the theorem, its meaning and its history along with examples appears in the text of Lad (1996, 99-120, 138-147). Computational applications to Chebyshev's inequality and Kolmogorov's inequality were presented by Lad, Dickey and Rahman (1992). Extended formulations of the theorem allow for the determination of bounds on conditional probabilities and for the implementation of constraining assertions that are non-linear in the argument variables, $\mathbf{q}_K$. Here we shall merely state the simplest form of the theorem, which should be sufficient to display that it provides the basis for the computational results that follow. After statement of the theorem, we shall remark on some generalisations. In Section 6, we shall remark further on some recent computational advances that are relevant to the problem we have proposed.

The Fundamental Theorem of Probability (FTP): Suppose $\mathbf{X}_N$ is any vector of events for which probabilities have been asserted, and let X_{N+1} be any other event of interest. Suppose further that there are K possible ovservations for the unknown vector $\mathbf{X}_{N+1}$. (Thus, the realm matrix $\mathbf{R}(\mathbf{X}_{N+1})$ has dimension $(N+1) \times K$. We partition it into its first N rows and the final row,

$$\mathbf{R}\begin{pmatrix} \mathbf{X}_N \\ X_{N+1} \end{pmatrix} \quad = \quad \begin{pmatrix} \mathbf{R}_{N,K} \\ \mathbf{r}_{N+1} \end{pmatrix} \tag{10}$$

because these parts of the realm matrix play different roles in the following.) Then the further assertion of a numerical value for $P(X_{N+1})$ coheres with the assertions specified by $P(\mathbf{X}_N)$ if and only if it lies within the interval $[L, U]$, where

$$\begin{array}{rcl} L & = & min\, \mathbf{r}_{N+1}\, \mathbf{q}_K \quad \text{and} \\ U & = & max\, \mathbf{r}_{N+1}\, \mathbf{q}_K \end{array}$$

subject to the conditions that $P(\mathbf{X}_N) \ = \mathbf{R}_{N,K}\ \mathbf{q}_K$ along with $\mathbf{1}_K^T\ \mathbf{q}_K = 1$, and $\mathbf{q}_K \geq \mathbf{0}_K$. The bold numbers $\mathbf{1}_K$ and $\mathbf{0}_K$ denote K dimensional vectors of 1's and of 0's, respectively.

The simplest generalisation of the FTP restates it as the fundamental theorem of prevision which applies to any unknown quantities, not merely events whose possible values are only 0 and 1. Applied to quantities whose realms exceed this limitation, the only difference is that the columns of $\mathbf{R}(\mathbf{X}_{N+1})$ list all possible vector values of the unknown quantity measurements, and thus will contain additionally numbers different from 0 and 1. De Finetti referred to both Probabilities and Expectations generically as "Previsions". Recall that in the standard theory of probability and expectation, a probability is a special case of expectation, the special case of the expectation of the indicator variable for the event whose probability is being assessed. Both are subsumed in the nomenclature of prevision, and designated by the same symbol, P. The statement of the theorem and its notation are identical, merely requiring the replacement of the words "probability" and "event" by "prevision" and "quantity".

Another difference of our specific problem from the simple statement of the FTP is that the probabilities we desire to assess with bounds are not unconditional but conditional probabilities such as $P(D^*|F)$ and $P(F|D^*)$. The text of Lad (1996) discusses the relevant form of the fundamental theorem, showing it to yield a fractional programming problem with a computable solution. Rather than discuss such details here, in Section 6 we shall refer to recent research that has greatly simplified computational procedures for bounds on conditional probabilities. It is still another matter in the problem we are considering that asserted inequalities on conditional probabilities specify quadratic constraints on $\mathbf{q}_{16}$. The linear restrictions that appear in the statement of the FTP need to be supplemented with an array of 15 quadratic inequalities, each in the form $\mathbf{q}_{16}^T \mathbf{A} \mathbf{q}_{16} \ \leq \mathbf{q}_{16}^T \mathbf{B} \mathbf{q}_{16}$. Quadratic restrictions provoke a second generalisation of the theorem.

We are ready now to report the results of our computations. Table 1 displays the computed bounds on a whole column of interesting probabilities that are restricted by coherence with the assertion of conditional exchangeability and the further inequalities we have discussed in Section 4. They are compared with the unique probabilities implied by the T-M formulation presuming conditional independence of the radiologists' judgments. We shall first outline the organisation of the rows of the Table, and then describe the differences in the several columns as we discuss them.

The various probabilities have been partitioned in Table 1 into blocks of self-contained interest. The first block pertains to probabilities characteristic of the accuracy of an individual assessor's diagnosis:

sensitivity, specificity, positive and negative predictive values. The next block presents these same characteristic probabilities as they pertain to the median diagnosis. Then the assessed probability of fibrosis for a patient in the presenting population appears alone. Block four assesses the differences between probabilities that were considered to be problematic in the inequality specifications of Section 4.4, along with two further probabilities of interest for the assessment of median diagnoses.

The first three pairs of numerical columns of Table 1 display comparisons of T-M's assessments based on conditional independence with our computations based on conditional exchangeability and the inequality constraints, the columns headed by "CondExFIC". Based on empirical studies which they cite, T-M had suggested an array of possibilities for the asserted value of individual sensitivity, p, running from .82 through .90 to .96. To begin the analysis of the results, it should be recognised how regularly the unique probabilities calculated according to the presumption of conditional independence do not fit within the intervals of probabilities that cohere with the conditions of CondExFIC for comparative columns. This is true both for the characteristics of individual diagnoses in block 1 and of the median diagnosis probabilities shown in block 2. This does not mean that the T-M probability estimates are incoherent in themselves. It is not incoherent to assert conditional independence. However, that presumption was used to make ballpark computations because of an inability of their analysis to deal with dependencies. It is notable that the resulting computations yield misleading estimates relative to the coherence condition with conditional exchangeability and the array of inequalities that appear reasonable.

Most notably, the presumption of conditional independence of the individual judgments overstates the positive gains to be expected from the median diagnosis procedure. Based on conditional exchangeability, the four fundamental probabilities of sensitivity, specificity, and positive and negative predictive values are expected to be greater for the median decision than for an individual decision, but the T-M approximation almost always exaggerates the expected gain. In particular, their assessment of the positive predictive value is well out of the order of what coherency requires based on what we can comfortably assert ... not only for the median diagnosis but for the individual diagnoses as well. To focus on one specific case, notice that when the individual sensitivity, $P(D_i|F)$, is specified as .82 then T-M identify the individual positive predictive value as .740 and the median value as .961. The corresponding intervals of probability induced by CondExFIC are $(.096, .685)$ and $(.146, .874)$, respectively.

Table 1: The first three pairs of columns compare probabilities of interest when computed according to Tweedie and Mengerson (T-M) and according to Conditional Exchangeability along with the Further Inequalities (CondExFIC) specified in Section 4.4. The next pair of columns further constrain the Median Decision probability PV+ at the lower and upper extremes of values that cohere with CondExFIC(.82). The final column is computed using CondExFIC along with a reasonable bound on PV+ which is discussed in the text. All computations presume $P(D^*) = .12$ and $P(S^*D^*) = .0504$. A star * by any other entry in the Table means that that entry was presumed as an input in the computation of any bound shown in that column.

Probability	T-M	CondExFIC	T-M	CondExFIC	T-M	CondExFIC	LoMedPV+	UpMedPV+	MedPV+*
$p = P(D_i\|F)$	.82*	.82*	.90*	.90*	.96*	.96*	.82*	.82*	.82*
$1 - p_f = P(\widetilde{D_i}\|\widetilde{F})$	.957	(.844, .951)	.893	(.845, .937)	.868	(.844, .921)	(.844, .878)	.951	(.905, .942)
$PV+_{ind} = P(F\|D_i)$	.740	(.096, .685)	.468	(.100, .576)	.374	(.091, .450)	(.096, .121)	.685	(.503, .626)
$PV-_{ind} = P(\widetilde{F}\|\widetilde{D_i})$	.973	(.975, .996)	.988	(.989, .998)	.996	(.996, 1.00)	(.995, .996)	.976	(.976, .980)
-----------	-----	-------	-----	-------	-----	-------	-------	-----	-------
$P(D^*\|F)$	.994	(.820, .907)	.972	(.900, .969)	.995	(.960, .995)	(.878, .890)	.906	(.871, .907)
$P(\widetilde{D^*}\|\widetilde{F})$	.995	(.895, .983)	.968	(.895, .964)	.952	(.894, .949)	(.895, .896)	.983	(.966,.970)
$PV+ = P(F\|D^*)$	.961	(.146, .874)	.761	(.149, .725)	.633	(.130, .599)	.14648*	.87312*	(.746,.771)*
$PV- = P(\widetilde{F}\|\widetilde{D^*})$	.988	(.981, .998)	.997	(.991, .999)	.9996	(.996 , 1.00)	(.997, .998)	.998	(.984,.990)
-----------	-----	-------	-----	-------	-----	-------	-------	-----	-------
$P(F)$	.126	(.019, .116)	.094	(.018, .091)	.076	(.015, .073)	(.019, .021)	.116	(.098, .107)
-----------	-----	-------	-----	-------	-----	-------	-------	-----	-------
$P(D_2\|\widetilde{D_1}F) - P(D_2\|\widetilde{F})$	.777	(.113, .727)	.777	(.139, .816)	.777	(.118, .864)	(.508, .589)	.727	(.522, .714)
$P(D_2\|D_1\widetilde{F}) - P(D_2\|F)$	-.777	(-.657, -.129)	-.777	(-.736, -.215)	-.777	(-.730, 0)	(-.292, -.129)	-.625	(-.657, -.446)
$P(D_3\|\widetilde{D_1}\widetilde{D_2}F) - P(D_2\|\widetilde{D_1}\widetilde{F})$	.777	(.010, .621)	.777	(.006, .761)	.777	(.003, .766)	(.316, .610)	.620	(.221, .619)
$P(D_3\|\widetilde{D_2}D_1F) - P(D_2\|F)$	0	(-.315, -.000)	0	(-.394, -.002)	0	(-.452, 0)	(-.090, -.029)	-.011	(-.115, -.009)
$P(D_3\|D_1\widetilde{D_2}F)$	.82	(.505, .820)	.90	(.506, .898)	.96	(.507, .960)	(.729, .791)	.809	(.705, .811)
$P(D_3\|D_1\widetilde{D_2}\widetilde{F})$	.043	(.103, .318)	.107	(.135, .327)	.132	(.180, .336)	(.186, .315)	.104	(.104, .188)
-----------	-----	-------	-----	-------	-----	-------	-------	-----	-------

This particular overstatement is especially notable because the unreasonably high value of .961 for PV+ for the T-M($p = .82$) computation is what motivated them to suggest that an individual sensitivity value of $p = .90$ is more reasonably supported by their derivations than is .82. It implies (for their computation) a value of .761 for PV+ which they assess as reasonable, and argue that their corresponding value of .633 when $p = .96$ is too low. The surprisingly large range of cohering values for the median PV+ relative to specifications of p when computed under CondExFIC presumptions weakens their argument dramatically. It further strengthens the assessment of .82 which was promoted by the large data study of Hughes and Weill (1991).

A second notable feature of the T-M proposal is the very low value of specificity for individual diagnosis which their argument implies but they did not publish. The value of $PV+_{ind} = .468$ shown in the column of T-M($p = .90$) would mean that an individual B-reader's classification of 1/0 or greater would support a higher probability for no fibrosis than for fibrosis. Although the difficulties of X-ray diagnosis of asbestosis are well-recognised, this is hardly believable. No physical reasons for the exhibition of small opacities in X-rays from patients without feruginous masses in the lung have been advanced. It is further interesting that the FTP allows rather large intervals of cohering sensitivity values for individual diagnosis, ranging from unbelievable low minimums under the full array of sensitivity specifications. We shall see that further considerations may allow us to narrow them.

Thirdly, it bears noticing that the cohering intervals for $PV+$ are the widest of all the computed cohering intervals that appear in the Table. It is useful to compare what the implications would be for the remaining probabilities if each of the extreme values of the positive predictive value for the median decision procedure were actually merited by evidence. The results appear in the seventh and eighth columns of intervals headed by "LoMedPV+" and "UpMedPV+". One remarkable feature of these computations is that at the higher specification of PV+ (.87312), the lower bounds and upper bounds on all cohering probabilities now agree through three decimal places, despite the fact that there are still three free dimensions in our uncertainty about the components of $\mathbf{q}_{16}$. The bounding intervals cohering with the lower PV+ value narrow markedly as well. However, the location of many of these intervals corresponding to "LoMedPV+" are well out of the order of values that would be considered sensible. Notice that the individual $PV+_{ind}$ would be bounded quite unreasonably within (.096, .121), and that the probability of fibrosis itself in a presenting population of at-risk-workers would be quite unreasonably bounded within (.019, .021).

These considerations motivate our computation of the final column of the Table, headed "MedPV+*". This results from a rather tight interval restriction on the median decision positive predictive value, asserting $PV+$ within the interval $(.746, .771)$. An interior value of $PV+ = .761$ was identified as an acceptable value by T-M, supporting their preferred value for individual sensitivity of $p = .90$. We can now see in the final column that appending a tight interval around .761 as the assertion PV+ accounts for both reasonable and fairly tight intervals of probabilities of interest under conditional exchangeability and the Hughes-Weill motivated value of $p = .82$. A worthy feature is that the median specificity is characterised as slightly larger than the sensitivity. This would be a fairly important characteristic of accuracy values for diagnosis procedures in situations as difficult as the X-ray diagnosis of asbestosis. Also of note is that the coherency restriction it forces upon $PV+_{ind}$ is $(.503, .626)$, a very reasonable interval.

The fourth bank of bounds displayed in Table 1 shows that several seemingly problematic probability orderings underlying the distinct inequality relations asserted in inequalities (4-7) are actually resolved on the basis of other assertions that have been made. Note firstly that $P(D_2|\widetilde{D_1}F)$ exceeds $P(D_2|\widetilde{F})$, though the extent of exceedence has a great range. Thus, one negative diagnosis on a patient with fibrosis is not sufficient to reduce the probability of the next positive diagnosis below that of a patient without fibrosis who has not yet been diagnosed at all. A similar result holds for the contrapositive, but shifted by a value of about .10 : $P(D_2|D_1\widetilde{F}) - P(D_2|F)$ is seen to exceed -.65, but could be as large as 0. Continuing through the block, we find that a second negative diagnosis on a patient who has fibrosis still motivates a higher probability of positive diagnosis from the third radiologist than does a single negative diagnosis on a patient without the condition. Although these bounds on differences are still wide, the direction is unequivocal. The closest of the differences to 0 occurs in the comparison of $P(D_3|\widetilde{D_2}D_1F)$ with $P(D_2|F)$. Although the latter is larger, they could be as much as equal.

The final two probability intervals pertain to events in which the first two readers disagree in their assessment of fibrosis. Looking only at column CondExFIC(.82) for example, it is interesting that while

$P(D_i|F) = .82$, $P(D_3|D_1\widetilde{D_2}F)$ is bounded within $(.505, .820)$. Hearing differing X-ray diagnostics for a patient with fibrosis reduces probability for the next positive diagnosis toward .5 but not quite to it. Similarly, while $P(D_i|\tilde{F})$ is within $(.049, .156)$, bounds for $P(D_3|D_1\widetilde{D_2}\tilde{F})$ are given by the interval $(.103, .318)$. Again the differing diagnoses push the cohering conditional probability for the next positive diagnosis toward .5, but it does not quite reach that balance point. This computation may be of wide substantive interest. One research oncologist with whom we consulted was adamant that differing X-ray diagnoses by two readers should motivate a probability of the next positive diagnosis as .5, irrespective of whether the patient actually has fibrosis condition or not. In the context of the CondExFIC(.82) assertions, such a further assertion would be incoherent.

6 Computational technique

While computational techniques of quadratic programming have been available for some time, it was surprising to us that the particulars of this problem did not yield directly to computation via the software available in MAPLE, a usually reliable programming code. We experienced computational problems that are well-known to quadratic and linear programmers, mainly cycling and erratic non-reproducable results from run to run depending on the condition of the machine. However, we were fortunate that the problems we addressed were small enough that we could actually compute the bounds in a direct way by running through a four-dimensional grid of feasible probabilities in the convex boundary implied by CondExFIC. Recall that the conditional exchangeability assertions reduce the dimension of the problem from sixteen to eight, and to seven when the unit summation constraint is imposed. The specifications of $P(D^*), P(S^*|D^*)$ and the individual sensitivity probability $P(D_i|F)$ further reduce the free dimensions to four. We iterated our computations in a grid through the simplex of linear constraints on the four remaining variables, and ignored any results that did not satisfy the fifteen quadratic inequalities and the two linear inequalities that we motivated in Section 4.4. The fineness of the grid was run on the order of 10^{-4} through each dimension. MAPLE code is available from the authors.

Having said this, it is worth reporting two very important advances that have been made in computational aspects of the types of problems we have applied here. One is in the formulation of problems that require the computation of bounds on a conditional probability. The text of Lad (1996, pp. 139-142) discusses a resolution based on the naive formulation of $P(A|B)$ as $P(AB)/P(B)$, reducing the objective function to the quotient of two linear functions of a vector such as $\mathbf{q}_{16}$ in our problem here. However, further analysis by Coletti and Scozzafava (1996) transforms the problem algebraically into one with a purely linear objective function, by means of a projection argument. The simplex bounded in (B, AB) space can be projected along rays from the origin onto the line defined by $B = 1$, and the programming problem is suitably specified as linear.

Technical details of further developments are based on many years of theoretical work by Coletti and Scozzafava who have summarised it all in one monograph (2002). The basis for their research has been de Finetti's seminal work on coherency of conditional probability assertions for which the conditioning event is either assessed as zero, or is restricted by coherency with other assertions to equal zero, or cannot be bounded away from zero on the basis of the limited assertions that can be made in a problem. In completely characterising the resolution of such problems via a theory of "layers of zeros" in a finitely additive context, their work has yielded practical solutions to the problem of computing bounds on such conditional probabilities. Procedures involve deleting specific columns from constraint matrices appropriate to the problem. All interested readers should be aware of this work which includes computational examples as well as theory and foundational discussion.

7 Substantive discussion and conclusions

There are myriad questions of interest that could be discussed, as our discussion of the computations has displayed. In light of the publication of the T-W results, two substantive conclusions from our work should be stressed. The first is to highlight the value that median diagnosis procedures are seen to provide in problems that are known to provoke differences in assessment across assessors. While T-M seem to have been originally intrigued by this question, they do not really assess its conclusion in their article, neglecting to present and compare the differences in predictive values achieved via individual and median diagnosis. Their presumption of independence of individual diagnoses has been shown to overstate the gains achievable from median diagnosis. This would have been expected were it formally recognised that different B-readers would bring a lot of shared information to their assessments of the same X-rays. While our analysis using conditional exchangeability properly recognises this shared information base, it also shows quite strikingly that median decisions of even three diagnosticians can provide useful gains in predictive probabilities in problems such as asbestosis where the natural difficulties and variabilities of X-ray assessment limit the power of individual diagnoses based on the most complete reasonable CondExFIC computation. Note, for example, the final column of Table 1 shows that accuracy rates of the median diagnosis still achieve useful gains beyond the accuracies of individual diagnoses. This should allay fears such as expressed by Walter and Irwig (1988, pp. 934-935) who rightly worried about the extent to which "correlated diagnosis errors" by readers would affect the results of assessments of median decisions.

Secondly must be mentioned a very strong and, we believe, unsupported conclusion to the T-W article, doubting the value of median X-ray diagnosis of asbestosis as an indicator of asbestosis related lung cancer. It has been widely recognised that in providing an environment for the initiation of cancer cells, the condition of exposure to asbestos dust is exacerbated by the habit of cigarette smoking. Physical mechanisms can be outlined that detail the relation between the accumulation of tars/nicotine and the presence of the ferruginous bodies that stimulate the development of cancer cells. Thus, the condition of asbestosis of the lung is distinct from the condition of lung cancer, but it is generally recognised as a provocative condition for the development of cancer. Hughes and Weill (1991) had provided empirical evidence based on a clinical study of 642 asbestos manufacturing workers, finding that excess lung cancers above baseline rates were largely composed of workers who had been radiologically diagnosed as having fibrosis. In a brief remark, T-W claim to reduce the number of subjects who did have fibrosis using the value of $P(F) = .094$ that they computed on the basis of their conditional independence presumptions when $p = .90$. Moreover, they claim (p. 237) that "other things being equal, it then follows that the percentage of excess lung cancers in this positively diagnosed $\geq 0/1$ group reduces to 55%, indicating that excess lung cancer is almost as likely to be associated with a nonopacity as an opacity." Our results indicate coherent bounds on $P(F)$ that easily cover a range supporting Hughes-Weill empirical findings of excess cancers among radiologically diagnosed subjects. Moreover, the details of the T-W computation, which were not described specifically in the publication, involve an error. Correct computations based even on their own arguments do not support their critique of the use of X-ray diagnosis as a warning against cancer.

A third substantive issue concerns the possible loss of important information when individual B-readers' classifications are collapsed into a specification of $\geq 0/1$ or not. Central to this entire diagnostic problem are questions about the extent to which one radiologist's diagnosis would be informative to us about the diagnosis of the next radiologist. In fact, variation among B-readers' classifications are not so very disparate when they do not agree precisely. Hughes and Weill (1991, p. 230) report that diagnoses by three readers' showed 94%, 94% and 88% of their categorisations were within one sub-category of the median of their readings. (Remember that categories of possible diagnosis are denominated by separated pairs such as 1/2, or 3/2, and so on.) It would appear that diagnostic distinction should be made between the assessments of an X-ray that is classed, say as 1/2, 2/1, 2/2 by three readers from one that is assessed as, say 1/0, 1/0, 1/1. Similarly, but supporting a different median decision, the assessments 0/1, 0/1, 1/0 and 0/1, 0/1, 1/2 could mean very different things. A word to the wise should be sufficient to be concerned about the loss of valuable information by reduction of costly information distinctions identified by the readers' exact assessments.

Finally, we should stress the gains that could be achieved by the statistical community through a clearer attention to two fundamental issues: the difference between conditional independence and conditional exchangeability; and the use of intervals to represent subjective assessments based on limited information. The assertion of conditional exchangeability, which we have found to be relevant here, has far wider applicability. It is relevant particularly to many problems that are improperly portrayed in Bayes' networks as situations of conditional independence. See, for example, the discussion in Lad (1999). Furthermore, when limitations on informational background cannot motivate a completely assessed probability distribution, de Finetti's fundamental theorem of prevision provides a computational procedure for using efficiently the assertions that we are able to make.

Acknowledgements

This research was supported by the Institute for Advanced Study, University of Bologna, where Frank Lad was a visiting professor of statistics during May and June, 2002, and by the University of Perugia Department of Mathematics for visits during 2002 and 2003.

8 REFERENCES

Black, M. and Craig, B. (2002) Estimating disease prevalence in the absence of a gold standard, *Statistics in Medicine*, **21**, 2653-2669.

Bruno, G. and Gilio, A. (1980) Applicazione del metodo del simplesso al teorema fondamentale per le probabilita nella concezione soggettivistica, *Statistica*, **40**, 337-344.

Coletti, G. and Scozzafava, R. (2002) *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, "Trends in Logic" Series.

Coletti, G. and Scozzafava, R. (1996) Characterization of coherent conditional probabilities as a tool for their assessment and extension, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **4**, 103-127.

de Finetti, B. (1937) La prévision, ses lois logiques, ses sources subjectives, *Annales de L'Institute Henri Poincaré*, **7**, pp. 1-68. H. Kyburg (tr.) Foresight, its logical laws, its subjective sources, in H. Kyburg and H. Smokler (eds.) *Studies in Subective Probability*, second edition, 1980, New York: Krieger.

de Finetti, B. (1967) Quelques conventions qui semblent utiles, *Revue Roumaine des Mathématiques Pures et Appliquée*, **12**, 1227-1233. Reprinted in translation by L.J. Savage, A useful notation, in B. de Finetti, *Probability, Statistics and Induction: the art of guessing*, 1972, New York: Wiley, pp. xviii-xxiv.

de Finetti, B. *Theory of Probability* (1974,1975) 2 volumes, A.F.M. Smith and A. Machi (trs.), New York: Wiley.

Geisser, S. and Johnson, W. (1992) Optimal administration of dual screening tests for detecting a characteristic with special reference to low prevalence diseases, *Biometrics*, **48**, 839-852.

Hughes, J. and Weill, H. (1991) Asbestosis as a precursor of asbestos related lung cancer: results of a prospective mortality study, *British Journal of Industrial Medicine*, **48**, 229-233.

Lad, F. (1996) *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley.

Lad, F. (1999) Assessing the foundations of Bayesian networks: a challenge to the principles and the practice. *Soft Computing*, **3**, 174-180.

Lad, F., Dickey, J.M., and Rahman, M.A. (1992) Numerical application of the fundamental theorem of prevision, *Journal of Statistical Computation and Simulation*, **40**, 131-151.

Martin, L. (2002) Asbestos lung disease: a primer for patients, physicians and lawyers, www.mtsinai.org/pulmonary/Asbestos/asbestos-questions.htm, 1-23.

Tweedie, R. and Mengersen, K. (1999) Calculating accuracy rates from multiple assessors with limited information, *The American Statistician*, **53**, 233-238.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*, London: Chapman and Hall Monographs on Statistics and Applied Probability, Number 42.

Walter, S.D. and Irwig, L.M. (1988) Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review, *Journal of Clinical Epidemiology*, **41**, 923-937.

Appendix: A Note on T-M's Identity, $g[p,\, p_f;\, P(D^*), P(S^*|D^*)] = 0$

T-M derive an implicit functional identity, $g(p, p_f, p^*, p_f^*, p^\#, p_f^\#) = 0$, that they express in their notation via the equation

$$p_{D^*}(p^\# - p_f^\#) + p_f^\# p^* \quad = \quad p_{S^*} p_{D^*}(p^* - p_f^*) + p_f^* p^\# \tag{11}$$

where their p_{D^*} is equivalent to what we have denoted by $P(D^*)$ and their p_{S^*} is equivalent to our $P(S^*|D^*)$. The symbols denoted by $p^*, p_f^*, p^\#$ and $p_f^\#$ would correspond to the probabilities we designate by $P(D^*|F), P(D^*|\widetilde{F}), P(S^*|F)$ and $P(S^*|\widetilde{F})$. However, because of their presumption of conditional independence among individual diagnoses given F and given $\widetilde{F}$, T-M replace these probabilities in equation (11) by $p^3 + 3p^2(1-p)$, $p_f^3 + 3p_f^2(1-p_f)$, $3p^2(1-p)$ and $3p_f^2(1-p_f)$, respectively. These substitutions reduce equation (11) for them to a complicated equation that is cubic in both p and p_f, and also involves the product term $p^3 p_f^3$. Values of p_{D^*} and p_{S^*} are sufficient to evaluate their equation numerically, yielding their "identity" in two variables, p and p_f, parameterised by specifications of $P(D^*)$ and $P(S^*|D^*)$. Thus, it has the form $g(p, p_f; p_{D^*}, p_{S^*}) = 0$. For any value of p, their equation can yield as many as three solutions for p_f, including more than one within the interval $[0, 1]$. Moreover, the complete expansion of equation (11) using these substitutions forms an equation between two ratios. T-M do not mention explicitly the situation that would make the denominator of the ratio equal to 0. In the case that $p = .82$, there are two solutions for p_f within $[0, 1]$. One of these would also make the denominator equal to 0. T-M ignore it.

It is by this method that T-M determine a complete distribution over all 16 possibilities in the general problem from a specification merely of values for $P(D^*)$ and $P(S^*|D^*)$ and a hypothetical value of p. It should be recognised that the "identity" they use to compute their results holds only under the supposition of conditional independence of the radiologists' assessments, a judgment that is not merited by the situation. Moreover, on the face of it, their solution does not really make sense. Among many other questions that could be asked, we ask "Could anyone believe that reasonable assumptions for the general problem we are considering would yield an exact value for $P(F)$, for example, from the mere specifications of $P(D^*), P(S^*|D^*)$ and $P(D_i|F)$?" We think not. The analysis based on conditional exchangeability does not require this presumption. That is why the results of Table 1 show intervals for the specification of $P(F)$ under each column headed by CondExFIC. We have continued working with T-M's numerical assessment of values for $P(D^*)$ and $P(S^*|D^*)$ in the main text of this paper.