APPLICATIONS OF DECISION ANALYSIS

IN DIAGNOSTIC RADIOLOGY

Toepassingen van de besliskunde in de Radiodiagnostiek

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus Professor Dr. C. J. Rijnvos
en volgens besluit van het College van Dekanen

De openbare verdediging zal plaatsvinden op
woensdag 25 oktober 1989 om 15.45 uur

door

Maria Gabriëlla Margaretha Hunink

geboren te Enschede

Promotie commissie

Promotor:	Professor Dr. J. Lubsen
Overige leden:	Professor C.B. Begg, Ph.D.
	Professor Dr. H.E. Schütte
	Professor Dr. Ir. J.H. van Bemmel

"Formal logic is based largely on statements of the kind: All X is Y. All Y is Z. Therefore all X is Z. (...) Medical facts are rarely suitable for this discipline. They are more of this kind - Some X is Y, but I'm not really sure of it. Possibly all Y may be Z (as Smith has claimed to show) but Brown affirms that no Y is ever the slightest Z, and Robinson too has strong views, but I find them hard to understand because he uses X to mean both Y or Z, and he calls Y what I call X."

(Richard Asher: Talking Sense)

Aan mijn ouders

CONTENTS

I. INTRODUCTION 1
Quantifying uncertainty: probability 1
Limitations of judgment under uncertainty and reasoning errors 2
Why use decision analysis? 4
Expected utility theory 5
Brief historical review 5
Decision analysis and protocols 6
Fields related to decision analysis 6
Application of decision analysis in medicine 7
Decision analytical methods in diagnostic radiology: outline of this thesis 8

II. THEORY of DECISION TREES 13
Decision trees - basic concepts 14
Markov models 22
Rates and probabilities 24
Modelling time-dependent transition probabilities in Markov processes 28
Modelling life expectancy 29
Modelling patient preferences 34
The axioms of expected utility theory and an alternative theory: prospect theory 42

III. RENOVASCULAR HYPERTENSION 47
Introduction 48
The structure of the problem 48
Summary of available data and assumed probabilities 49
Assignment of utilities 55
Results 56
Comment 60

IV. PERCUTANEOUS NEPHROSTOMY FOR ACUTE URINARY TRACT OBSTRUCTION CAUSED BY UROLITHIASIS 65
Introduction 66
The problem stated 68
Summary of available data 69
The model 89
Technical details 91
Assumptions 92
Results 93
Discussion 106

V. THEORY of ROC ANALYSIS 113
Bayesian thinking and likelihood ratios 114
What is an ROC curve? 116
Sensitivity, specificity and ROC curves 117
Obtaining data for an ROC analysis 120
ROC curve models and indices 122
Comparing ROC curves 123
The optimal operating point 123
Workup bias or verification bias 125
Uninterpretability bias 136
VI. CT AND MRI ASSESSMENT OF ENT TUMOR EXTENSION USING ROC METHODOLOGY 139
Introduction 140
Material and methods 141
Results 145
Discussion 150
VII. ROC ANALYSIS OF THE CLINICAL, CT AND MRI DIAGNOSIS OF ORBITAL SPACE-OCCUPYING LESIONS 153
Introduction 154
Material and Methods 154
Results 161
Discussion 169
VIII. TESTING FOR FETAL PULMONARY MATURITY: AN ROC ANALYSIS INVOLVING COVARIATES, VERIFICATION BIAS AND COMBINATION TESTING 173
Introduction 174
Methods 176
Results 181
Discussion 188
IX. DISCUSSION 195
Decision analysis 196
ROC analysis 209
Conclusions 213
SUMMARY 217
NEDERLANDSE SAMENVATTING 220
CURRICULUM VITAE 223
ACKNOWLEDGEMENTS 225
ABBREVIATIONS 227
INDEX 229


Centrum Voor Klinische Besliskunde
ROTTERDAM
Center For Clinical Decision Analysis

Chapter I

INTRODUCTION

I. Quantifying uncertainty: probability
II. Limitations of judgment under uncertainty and reasoning errors
III. Why use decision analysis?
IV. Expected utility theory
V. Brief historical review
VI. Decision analysis and protocols
VII. Fields related to decision analysis
VIII. Application of decision analysis in medicine
IX. Decision analytical methods in diagnostic radiology: outline of this thesis
X. References

Decision making in medicine frequently involves coping with uncertainty. Uncertainty exists about the diagnosis, the consequences of the disease and the risks and benefits of further workup and treatment. Clinical decision analysis is an explicit technique for making choices in the face of uncertainty. It makes explicit the possible consequences and outcomes of the various options, as well as the likelihood of the outcomes and their values. Even though we often do not know everything with certainty, "the decision **has** to be made" (23) and the goal of decision analysis is to help in making that decision.

I. Quantifying uncertainty: probability

Initial clinical judgment is based on a patient's presenting signs and symptoms. Based on training and experience, the physician will have a sense of the likelihood that the patient has a particular disease given those signs and symptoms. The physician will develop a list of differential diagnoses, with the most likely disease at the top, and considers which tests might help in distinguishing among the possibilities. The next step is to adjust the initial clinical impression with the information obtained from the diagnostic tests. But diagnostic tests are seldom perfect and the possibility of a false positive or false negative test result has to be kept in mind. Again the physician will question him/herself as to how often a false positive or false negative test result occurs in the situation at hand. Lastly, treatment options and their likelihood of success and complications are assessed.

When considering the above questions, physicians often speak in terms of "probably", "very likely" or "can't be excluded". Different physicians may have different interpretations for the same probabilistic term, for example, likely may mean 60% for one and 75% for another (12). Thus, quantifying uncertainty forces physicians to be more precise. Uncertainty is quantified by expressing it as a probability. In decision analysis we assume that the "relative frequency" can be used to estimate the "probability". The notion is: if I had 100 patients identical to the patient sitting in front of me, how many would have the disease? If, for example, 20 would have the disease, then the probability of the disease in this patient would be 0.2[1]. The concept "probability" is fundamental to decision analysis. Thinking in terms of probabilistic information distinguishes clinical decision analysis from the more familiar deductive reasoning in medicine, that is, cause-effect type of reasoning.

II. Limitations of judgment under uncertainty and reasoning errors

Physicians are subject to the same cognitive errors and limitations that affect all human reasoning. They are subject to bias when estimating the probability of events or disease and have difficulty integrating probabilistic information (1,2,10,35). People use heuristics, that is "rules of thumb", to estimate the probability of an event, which may be useful, but, however, may also lead to incorrect estimates. Heuristics can be categorized as representativeness, availability, anchoring and adjustment heuristics (35).

Using the representativeness heuristic, the probability that A belongs to group B is judged by how representative A is of group B (35). However, by relying on representativeness to estimate probability we may neglect other important information or make elementary statistical errors, leading to biased probability estimates. Physicians may not take into account the prior probability[2] of disease appropriately (8), eg. a clinician instituted an extensive workup for a patient with hypertension and a positive phentolamine test and was surprised that a phaeochromocytoma was not found, forgetting the low prior probability of this type of tumor (10). We expect a short sequence of events to be representative of the whole process, which is not necessarily so, as shown by the gambler's fallacy: "if it has been tails many times in sequence, it is bound to be heads next" (this is not true: the probability that it will be heads equals 0.50, provided a fair coin is being tossed) (35). We do not intuitively infer from experience principles such as regression towards the mean , eg. we are surprised that in teaching a resident, positive reinforcement is

[1]A probability is always larger or equal to 0, and smaller or equal to 1. "Never" is equivalent to a probability of 0 and "always" is equivalent to a probability of 1.

[2]The prior probability of disease is the relative frequency of the disease in the population that the patient belongs to. The population may be defined by clinical features such as age, sex, race, signs and symptoms.

often followed by worse performance (35). The effect of sample size on sampling variability escapes our attention, that is, we forget that a small sample is more likely to give an incorrect estimate of a probability than a large sample, eg. a ratio of 3:2 of male to female newborns, is more likely to occur in a small community than in a large community, although we tend to think it should be 1:1 in both (35). Representativeness is usually considered equivalent to predictability, eg. "the case we had last month was just like this case" (as if what happened to last month's case will have any value in predicting what will happen to this case !).

Availability is appraised as equivalent to frequency, that is, the more cases that spring to mind the higher we judge the frequency to be (35). Instances of greater impact (such as seeing children with brain tumors in neuroradiology) and recent occurrences, cause us to overestimate the probability of that occurrence. Risks involved in doing a procedure may be underestimated if one has difficulty in imagining the possible dangers or if one neglects to think of these, eg. the inexperienced interventional minded radiology resident doing an invasive procedure may underestimate the risk involved.

People judging a probability with the anchoring and adjustment heuristic start from an initial value and then adjust the value to derive the estimate (35). However, the adjustments tend to be insufficient. Furthermore, when asked to give a range of possible estimates we tend to give a narrow confidence interval, reflecting more confidence in our estimate than we ought to have based on our information (35).

Apart from the inaccuracies that may occur when relying on heuristics, other factors may contribute to biased probability estimates, such as ego bias, hindsight bias and regret (1). Ego bias occurs when a probability estimate is distorted in a self-serving manner (1). For example, a radiologist might think his/her probability of success in performing an angioplasty higher than the reported average. Hindsight bias occurs frequently in medicine: after the diagnosis has been made, we tend to overestimate the probability we would assign to the diagnosis were we to be confronted with the problem anew. After an event has occurred we tend to say we had predicted it, or could have predicted it, beforehand (1). Regret may influence our probability estimates in that if an event has an undesirable outcome, we may overestimate the probability of its occurrence(1).

Apart from the inaccuracy of our probability estimates, we are unable to intuitively combine numerous facts (2,20). Humans can only consider five to nine pieces of information at one given time (20). A number of the facts will simply be ignored: this is a necessary process when confronted with an overwhelming amount of information but will not invariably provide us with the optimal decision. Furthermore, common components of various options are disregarded and we focus on the differences of the options: this simplifies the decision making problem but is not necessarily the best thing to do (9,36).

III. Why use decision analysis?

Decision analysis is an explicit way of decision making as opposed to the intuitive decision making familiar to us all. Most of decision making in medicine is based on uncertain facts: there is a certain probability that the patient actually has the disease and a probability that if we intervene he might benefit from it, or die as a result of our efforts. Risks and benefits constantly have to be weighed and balanced, with respect to both the probabilities that the events will occur and to the outcomes. Doctors are daily integrating information from their training, books, articles and experience to come up with what seems to them the most appropriate course of action (8). However, integrating information implicitly is associated with the limitations of judgment and reasoning errors discussed above. Furthermore, explaining the logic of implicit decision making to others concerned is difficult, if not impossible.

The limitations of judgment discussed above motivate the use of decision analysis. However, the same arguments are used against clinical decision analysis. Analysts sometimes rely, at least in part, on subjective probability estimates if objective data are unavailable and, thus, the decision analytical solution may be prone to the same biases as intuitive decision making. However, this usually involves only a limited number of probabilities and a method exists to test how sensitive the decision is to the particular piece of information, namely sensitivity analysis. Similar to intuitive decision making, modelling a decision often necessitates omitting less important information. Without simplifications, the decision tree and calculations may be become intractable. However, the amount of information that can be included in a decision analysis is far greater than the amount we can remember and integrate in our heads.

Doctors tend to defend their decisions on the grounds of experience. But what is experience? It is an implicit integration of frequencies, probabilities and outcomes as seen by that one physician and therefore limited to his/her experience and subject to bias and distortion.

Clinical decision analysis helps with the difficult task of decision making under uncertainty. It makes explicit the information used in the decision and indicates the optimal decision, based on the likelihood of different outcomes and their values. It helps us structure and analyze the overwhelming amount of medical information that exists and may be used as a means of discussing the problem with those concerned. The consequences of different assumptions and likelihood of events can be examined. Decision analysis does not always produce a definitive solution to the problem. However, it gives insight into the decision, identifies the trade-offs involved and information needed, and provides a consistency check of intuitive decision making.

IV. Expected utility theory

Decision analysis is based on expected utility theory (9,37). The utility of an outcome is the value associated with that outcome. The utility can be expressed as a gain or as a loss. A disutility refers to a utility expressed as a loss. The overall expected utility of a decision equals the sum of all the possible outcomes, each outcome weighted for the probability that it will occur. Folding back and averaging out refers to the arithmetic of multiplying each outcome by the probability that it will occur and adding the results, giving the expected utility of the option. The main axiom of expected utility theory is that an individual wishes to maximize expected utility if outcomes are expressed as gains, or equivalently, minimize expected disutility if outcomes are expressed as losses.

The utility, or value of an outcome, may be expressed on a single-attribute scale, such as fractional survival or lifes saved, or it may be on a multi-attribute scale, such as lifes saved, hospitalization days averted and dollars spent. In the latter case the utility of the various attributes must be combined in some fashion, so that the scale represents the values of the outcomes, taking the various attributes into account (34).

V. Brief historical review

The methods used in clinical decision analysis originate from operations research, game theory and mathematical decision analysis.

One of the first articles about medical decision making, written by Lusted (a radiologist) and Ledley (a mathematician), was published in 1959 (13) in Science. Raiffa wrote "Decision analysis: Introductory lectures on choices under uncertainty" in the late 60's (25), one of the first books about decision theory. In 1971 the article "Decision-making studies in patient management", written by Lusted, appeared in a leading medical journal, the New England Journal of Medicine (14). Weinstein and Fineberg's "Clinical Decision Analysis" (37) published in 1980 has become a standard textbook for medical decision theory. A more recent book is that by Sox et al (28). Nowadays, books on epidemiology or medical statistics usually contain a chapter explaining decision analysis. Furthermore, medical schools are including decision analysis in their curriculum, or have been advised to do so. Recently, leading journals have published review articles describing the applications of decision analysis and the progress that is being made (5,11,21).

Decision analysis is slowly having an impact on clinical reasoning. Initial efforts to analyze individual case problems showed that performing a good analysis takes so much time that the decision is often made before the results of the analysis are clear (24). However, a brief analysis

("quick and dirty analysis") gives insight into the problem and may solve it. Progress has been made with the analysis of commonly recurring clinical problems, in which case the analysis is applicable to a group of patients, and the results of the analysis may influence clinical practice.

The advent of the computer age probably led, in part, to the development of clinical decision analysis, because of the associated logical reasoning and easy access to information. With the advances made in computer technology, decision analysis is becoming easier to perform. Hospital information systems make it possible to obtain relevant data. The literature can quickly be reviewed with a medical database, such as Medline (19). Mathematics software, such as MathCAD (15) and spreadsheets (32), facilitate modelling of data. Software specialized in decision analysis, such as Decision Maker (3) and Supertree (31), free the analyst from the calculational burden of analyzing decision trees and because trees can be saved and modified, they encourage experimentation with alternative models.

VI. Decision analysis and protocols

A protocol should be distinguished from a decision analysis. Protocols are algorithms, that is a flow chart of the steps to be taken given a particular problem. Protocols can vary from simple rules of thumb to extensive flow charts with multiple subdivisions. Examples are found in many standard textbooks (6,18).

A protocol can be written as a report of a consensus meeting, defined by an authority on the subject, or simply defined by the policy maker of the ward. Usually in a protocol a number of points exist at which the optimal decision for further workup or treatment is uncertain. Decision analysis can be used to decide which workup or treatment strategy will be optimal. Furthermore, a protocol may be valid for most cases, but not necessarily for the individual patient. For individuals decision analysis may be used to decide on the optimal management.

VII. Fields related to decision analysis

Fields related to decision analysis are clinical epidemiology, biostatistics, artificial intelligence and technology assessment.

Clinical epidemiology is concerned with quantifying the frequency of disease, the prognosis and the effects of treatment as a function of the determinants (26) and plays a major role in the methodology of clinical trials. Biostatistics is concerned with the statistical analysis of medical research data. Both clinical epidemiology and biostatistics are basic medical sciences important to clinical decision analysis.

Artificial intelligence programs, or expert systems, are designed to aid in diagnosis and management and attempt to emulate a clinical expert. Among the many expert systems developed, a large number are rule-based computer systems (27,33), in which "rule" refers to an "if-then" statement: "if" identifies a situation based on signs, symptoms and/or laboratory results, and "then" specifies the diagnosis or management. An example of a rule-based system is MYCIN (27,33). More recent efforts to emulate clinical reasoning concentrate on programs organized around models of disease (33). These programs consist of hierarchical structures, including clinical states and the corresponding pathophysiological states on a more detailed level. INTERNIST is an example of such a program (27,33). Expert systems are useful as a decision making aid, as a database, or as an alternative to a textbook. Furthermore, the development of expert systems has contributed to our understanding of medical reasoning.

Technology assessment is a comprehensive analysis of a technology, such as drugs, diagnostic tests, procedures or health care delivery systems. The assessment ideally encompasses technical, clinical, economic, social and ethical issues and may include activities as data acquisition, decision or cost-effectiveness analysis, synthesis of information (meta-analysis) and consensus meetings (22,30,38). Technology assessment is usually coordinated by government-linked bodies, research councils, advisory bodies or professional groups (30,38).

VIII. Application of decision analysis in medicine

The application of clinical decision analysis takes place on three different levels:

1. the individual patient
2. a group of similar patients
3. public health in general

Of course, some examples fit into more than one category, but this classification helps identify the scope of decision analytical models.

1. The individual patient

Decision analyses for individual patients have been done for decisions involving uncertainty about the diagnosis, uncertainty about treatment efficacy or treatment risks, increased risk of a test or treatment, limited benefit of therapy, competing risks and benefits, uncertainty about the optimal timing or optimal sequencing of procedures, explicit patient preferences, uncertain medical information, or involving a rare, unique, or new problem (24). A clinical decision consultation service has been offered to clinicians since 1978 by the Division of Clinical Decision Making at the New England Medical Center (Tufts University) in Boston (24). Many of the more interesting problems analyzed at the Division have been published in the form of "Clinical

Decision Making Rounds" in the journal "Medical Decision Making". More recently, the Center for Clinical Decision Analysis at the Erasmus University and Dijkzigt University Hospital in Rotterdam has started a similar service.

2. A group of similar patients

Generic decision models typically address common clinical problems. It is useful to perform the analysis for different types of patients, ie. for different age, sex, disease severity and/or risk groups. The model should be extensively tested and reviewed before it is used to give general recommendations for standard practice. An example is the indication for coronary angiography in patients with chest pain (4).

3. Health policy problems

Health policy problems often concern screening or vaccination programs, such as screening for hypertension (17,29), and usually require a cost-effectiveness or cost-benefit analysis. A cost-effectiveness analysis examines financial costs in relationship to the expected benefit of a program, expressing the results as the ratio of marginal (monetary) costs to the marginal effectiveness (37). A cost-benefit analysis also examines financial costs in relationship to the expected benefit of a program, but expresses the results in monetary value which necessitates valuing life in dollars (or another currency) (37).

IX. Decision analytical methods in diagnostic radiology: outline of this thesis

A number of radiologists have made important contributions to the field of medical decision making (5,14,16). The two most useful decision analytical techniques in diagnostic radiology are:

1. decision trees and Markov processes, especially concerning generic problems
2. receiver operating characteristic (ROC) methodology

Besides the two decision analytical methods mentioned various other techniques are used in radiology to evaluate data and facilitate decision making. Examples of such techniques are logistic regression analysis, discriminant analysis, and Bayes theorem (5).

The number of papers applying decision analytical tools to problems related to diagnostic radiology is slowly increasing. Receiver operating characteristic (ROC) analysis is probably the most extensively applied tool in diagnostic radiology, and used more and more often to evaluate and compare diagnostic tests. However, a number of methodological issues related to ROC

analysis are now becoming evident. Some of these issues are discussed in this thesis. The use of decision trees and Markov processes to examine common clinical problems has only been done sparingly. Examples of both types of models, applied to radiological problems, are presented in this thesis.

The subjects of this thesis are decision analysis and receiver operating characteristic (ROC) methodology applied to radiological problems. This thesis is intended for those interested in applying decision analytical techniques in diagnostic radiology, and in medicine in general.

Chapter II deals with the theory of decision trees and Markov processes. The basic concepts are briefly explained and a few selected topics are discussed in more detail. Chapter III describes a decision model for the diagnostic workup and treatment of renovascular hypertension. Chapter IV presents a Markov analysis of the decision whether, and when, to intervene in acute urinary tract obstruction. Chapter V deals with the theory of receiver operating characteristic (ROC) methodology. Basic concepts are explained and a number of selected issues are discussed in detail. Chapter VI presents an ROC analysis of the assessment of tumor extension in neoplastic disease of the nose, paranasal sinuses, nasopharynx and parapharyngeal space, comparing computer tomography (CT) and magnetic resonance imaging (MRI). Chapter VII presents the results of an ROC analysis of orbital space-occupying lesions comparing the diagnosis made by means of clinical evaluation, computer tomography (CT) and magnetic resonance imaging (MRI). Chapter VIII presents the results of a study on fetal pulmonary maturity testing, involving a number of interesting methodological issues of ROC analysis. A general discussion follows in chapter IX.

X. REFERENCES

1. Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. Journal of General Internal Medicine 1987; 2: 183-187.

2. de Dombal FT. Picking the best tests in acute abdominal pain. Journal of Royal College of Physicians of London, 1979; 13/4: 203-208.

3. Decision Maker. Division of Clinical Decision Making, New England Medical Center, Boston, Massachussetts.

4. Doubilet P, McNeil BJ, Weinstein MC. The decision concerning coronary angiography in patients with chest pain: a cost-effectiveness analysis. Med Decis Making 1986; 5/3: 293-309.

5. Doubilet PM. Statistical Techniques for Medical Decision Making: Applications to Diagnostic Radiology (Review). AJR 1988; 150: 745-750.

6. Eisenberg RL (ed) Diagnostic Imaging: an algorithmic approach. 1988. JB Lippincott Company, Philadelphia.

7. Halvorsen KT. Combining results from independent investigations: meta-analysis in medical research. In: Medical uses of statistics. Bailar III JC, Mosteller F (eds). 1986. N Engl J Med Books, Waltham, Massachusetts.

8. Henrik R Wulff. Rational Diagnosis and Treatment: an introduction to clinical decision making. 2 ed, 1981. Blackwell Scientific Publications, Oxford.

9. Kahneman D and Tversky A. Prospect theory: an analysis of decision under risk. Econometrica 1979; 47/2: 263-291.

10. Kassirer JP, Kopelman RI. Cognitive errors in diagnosis: instantiation, classification, and consequences. Am J Med 1989; 86: 433-441.

11. Kassirer JP, Moskowitz AJ, Lau J, Pauker SG. Decision analysis: a progress report. Ann Intern Med 1987; 106: 275-291.

12. Kong A, Barnett O, Mosteller F, Youtz C. How medical professionals evaluate expressions of probability. N Engl J Med 1986; 315: 740-744.

13. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959; 130: 9-21.

14. Lusted LB. Decision-making studies in patient management. N Engl J Med 1971; 284: 416-424.

15. MathCAD. MathSoft Inc., Cambridge, Massachussetts.

16. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975; 293: 211-215.

17. McNeil BJ, Varady PD, Burrows BA, Adelstein SJ. Cost-effectiveness calculations in the diagnosis and treatment of hypertensive renovascular disease. N Engl J Med 1975: 293: 216-221.

18. McNeil BJ and Abrams HL (eds). Brigham and Women's Hospital Handbook of Diagnostic Imaging. 1986. Little, Brown and Company, Boston/Toronto.

19. Medline. MDTV, de Medicus, Leiderdorp, The Netherlands.

20. Miller GA. The magical number seven, plus or minus two. Psychol Rev 1956; 63: 81-97.

21. Pauker SG, Kassirer JP. Medical Progress. Decision analysis. New Engl J Med 1987; 316: 250-258.

22. Pauker SG. Decision analysis as a synthetic tool for achieving consensus in technology assessment. International Journal of Technology Assessment in Health Care 1986; 2/1: 83-97.

23. Personal communication Milton Weinstein, quotation.

24. Plante DA, Kassirer JP, Zarin DA, Pauker SG. Clinical decision consultation service. Am J Med 1986; 80: 1169-1176.

25. Raiffa H. Decision analysis: Introductory lectures on choices under uncertainty. 1 ed 1968, Random House, New York.

26. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Little, Brown and Company.

27. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine. Where do we stand? N Engl J Med 1987; 316: 685-688.

28. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Butterworth Publishers, Boston, 1988.

29. Stason WB, Weinstein MC. Allocation of resources to manage hypertension. N Engl J Med 1977; 296: 732-739.

30. Stocking B. Strategies for technology assessment and implementation in some European Countries. International Journal of Technology Assessment in Health Care 1986; 2/1: 19-26.

31. Supertree. SDG Decision Systems, Menlo Park, California.

32. Symphony. Lotus Development Corporation, Cambridge, Massachussetts.

33. Szolovits P, Patil RS, Schwartz WB. Artificial intelligence in medical diagnosis. Ann Intern Med 1988; 108: 80-87.

34. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ 1986; 5: 1-30.

35. Tversky A and Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science 1974; 185: 1124-1131.

36. Tversky A and Kahneman D. The Framing of Decisions and the Psychology of Choice. Science 1981; 211: 453-458.

37. Weinstein MC and Fineberg HV. Clinical Decision Analysis. 1980. Saunders WB Co. Philadelphia.

38. Young DA. Strategies in technology assessment and implementation in the United States. International Journal of Technology Assessment in Health Care 1986; 2/1: 13-18.

Chapter II

THEORY of DECISION ANALYSIS

I. DECISION TREES - BASIC CONCEPTS
1. Structuring the decision tree
2. Assigning probabilities
3. Outcome values / utility assessment
4. Expected utility and folding back
5. Sensitivity analysis

II. MARKOV MODELS

III. RATES AND PROBABILITIES
1. The prior probability and prevalence
2. The cumulative incidence or cumulative failure function
3. The incidence density, hazard function or hazard rate

IV. MODELLING TIME-DEPENDENT TRANSITION PROBABILITIES IN MARKOV PROCESSES

V. MODELLING LIFE EXPECTANCY
1. The declining exponential approximation of life expectancy: DEALE-method
2. Disease specific excess mortality rates
3. The Gompertz model
4. The Weibull model

VI. MODELLING PATIENT PREFERENCES
1. Quality of life with morbidity
2. Risk aversion and risk seeking attitudes

VII. THE AXIOMS OF EXPECTED UTILITY THEORY AND AN ALTERNATIVE THEORY: PROSPECT THEORY

VIII. REFERENCES

The basic theory of decision analysis is extensively explained in Weinstein and Feinberg's textbook "Clinical Decision Analysis" and in the more recent textbook by Sox et. al. "Medical Decision Making". Sections I and II of this chapter briefly summarize the basic concepts of decision analysis. For further detail on the basic concepts I refer to the standard textbooks and the references. Section III reviews concepts from clinical epidemiology which, although basic to decision analysis, are not included in textbooks on decision analysis. In section IV an equation is derived for calculating time-dependent probabilities in Markov processes. I have included this because the derivation is neither immediately obvious nor have I found it in the literature. Section V explains three commonly used models of life expectancy. Section V.1 explains the

DEALE-method as introduced by Beck et.al. Section V.2 discusses two methods for calculating excess mortality rates. I have included this explanation because the calculation of excess mortality rates from survival data is often done incorrectly when applying the DEALE-method. Furthermore, although practical, the second method is not commonly described or used. In section VI.1 patient preferences towards quality of life and the existing approaches to elicit preferences are discussed. A new unit is introduced, namely PALY's (preference-adjusted life years). In section VI.2 attitudes towards risk are explained and a new type of utility curve, the "intern's utility curve" is introduced. Section VII discusses the axioms of expected utility theory.

I. DECISION TREES - BASIC CONCEPTS

Decision analysis uses decision trees to represent clinical strategies and their expected outcomes. Decision trees help clarify and delineate complicated problems by explicitly structuring the many probabilistic events and outcomes. However, seemingly clear-cut decisions may also benefit from structuring the relevant information in the form of a decision model. To illustrate the basic concepts of decision trees, consider the following example.

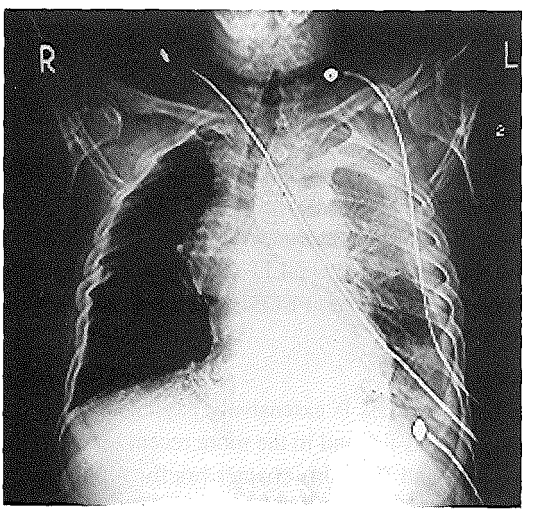


Figure 1. Chest X-ray of a car accident victim suggesting thoracic aortic rupture. The image shows a large apical cap on the left, widening of the superior mediastinum, deviation of the trachea to the right and poor delineation of the aorta.

A car accident victim, suffering from deceleration trauma, is brought to the emergency ward. His chest X-ray suggests thoracic aortic rupture (figure 1). The surgeon on call requests a CT scan of the chest, with contrast, to determine if the rupture is proximal or distal to the left subclavian artery. If the rupture is proximal to the left subclavian, a heart-lung-bypass machine would be necessary to perform surgery. The latter is not available at the hospital where the patient is admitted, so he would have to be transferred to another hospital if the rupture is proximal. The radiologist, however, would prefer not to perform the CT, because the patient might die while the examination is being performed. The problem stated: should a CT be performed or should the patient go to surgery immediately, presuming that the rupture is distal to the left subclavian artery?

1. Structuring the decision tree

A decision tree (figure 2) represents potential clinical strategies, and their consequences (21,25). Time proceeds chronologically from left to right. The tree starts, at the left, with a decision node , customarily depicted with a square. This node represents the point in time at which the decision is to be made. The branches of the decision node delineate the possible clinical strategies, under the control of the decision maker, for the particular problem, in this case CT-scan or immediate surgery. Chance nodes, represented by circles, represent uncertain states of health or events that result from the initial decision. The numbers underneath each branch represent the probability of that event. All probabilities at chance nodes must sum to one. In the example, the modelled events are death due to delay of surgery, the presence of a proximal versus distal rupture and death due to transfer to another hospital. At the end of each path of the tree is a terminal node which represents the outcome, or endpoint, of the events and disease states that precede it. In this example, the outcomes are dead or alive. Dead is assigned a value of zero and alive a value of one. Label nodes are sometimes interposed to arrange the tree conveniently and clearly, but they have otherwise no function. Decision nodes may be implied if the preceding event determines the decision, for example "transfer" is an implied decision node. Another type of node is a Boolean node, that is, a node at which a logical variable determines which path is followed. For example, in the above model, a Boolean node involving blood pressure as logical variable could have been included. If the blood pressure were greater than a specified amount, then a delay may be acceptable, although still perhaps a risky strategy.

For illustrative purposes this simple example does not include many events such as surgical mortality, allergy to contrast and the test characteristics (sensitivity and specificity) of CT scan in diagnosing thoracic aortic rupture. A decision model is always a compromise between simplicity and reality. Essential is that the trade-off of the clinical decision problem is modelled. In the example the trade-off is the information obtained from doing the CT scan versus the risk of death due to delay of surgery. Every model is based on a number of assumptions to simplify the problem, which must always be stated explicitly. In this example the assumptions are:

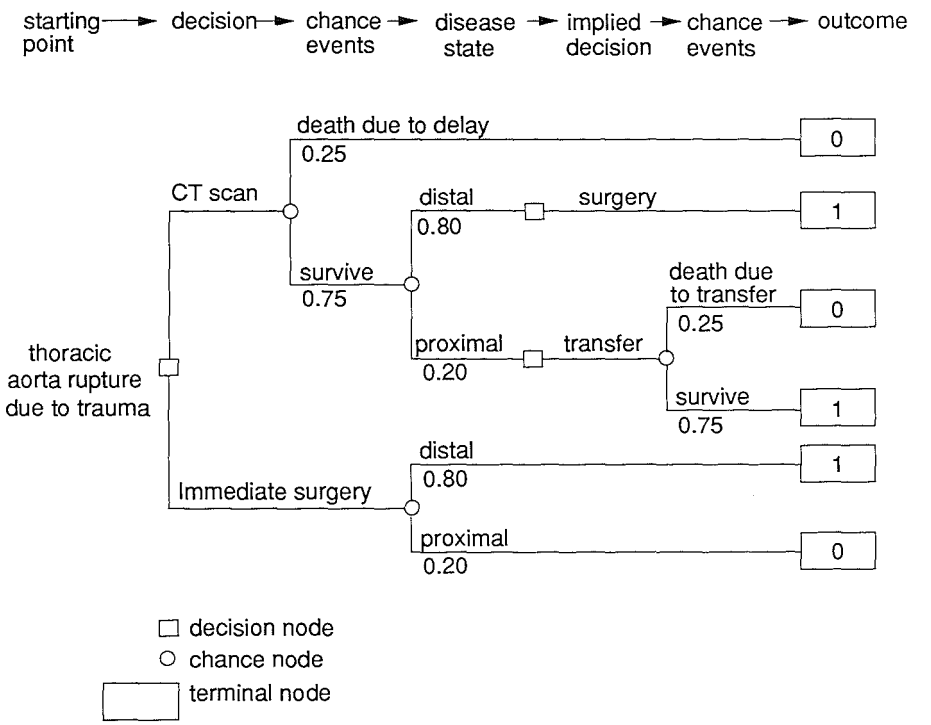


Figure 2. A simple decision tree to illustrate the basic concepts. Modelled is the trade-off between obtaining information by performing a CT scan and the probability of death due to the delay.

1) Performing a CT scan takes about half an hour, including transport to and from the CT unit. Transferring the patient to another hospital also takes about half an hour.

2) Patients who have a rupture distal to the left subclavian artery who are operated on without delay (in the hospital where admitted), all survive. Patients with a proximal rupture who are operated immediately without heart-lung-bypass machine (in the hospital where admitted), all die. Patients with a proximal rupture who are operated on with a heart-lung-bypass machine (in another hospital), survive.

Setting up a tree with binary chance nodes, common subtrees and similar modelling for each strategy, helps avoid errors in the model. Furthermore, it is advisable to model trees in strategic form, that is, there should be only one decision node, namely the one up front, and no decision nodes downstream, again to avoid errors. Modelling in strategic form can be accomplished using Boolean nodes and implied decision nodes. At a chance node the depicted chance events should

be mutually exclusive. One should define the time horizon of the analysis, that is, define clearly until what point in time the problem is modelled. As mentioned before, modelling a problem involves some trade-off between reality and simplicity and this should be kept in mind.

2. Assigning probabilities

Probabilities of chance events in the decision tree are derived by reviewing the literature, gathering data from a hospital information system or asking experts in the field to estimate the probability based on their experience (21,25). As mentioned in chapter I, we assume that the relative frequency of an event among patients similar to the patient under consideration can be used to estimate the probability of the event in the case patient. Clinical judgment, common sense and knowledge of statistics and study design are needed to decide whether a reported relative frequency is applicable to the case at hand. Occasionally a certain amount of manipulation of the available data is needed so that it can be applied to the decision problem (see chapter IV).

Considering the case example, 80% of patients with a traumatic aortic rupture have a rupture distal to the left subclavian artery (6), and therefore we set the probability for this patient at 0.80. For this example, we assume that the probability of death due to half an hour delay is 0.25, in other words, one-quarter of similar patients will die if surgery is delayed half an hour, and of the remaining patients another quarter will die if surgery is delayed an additional half hour.

3. Outcome values / utility assessment

Outcomes have to be valued in some quantitative way to perform the analysis. The utility of an outcome is the value assigned to the outcome (21,25). The simplest method to value outcomes is using a binary approach, either the patient is alive or dead, as done in the example. Alternatively, one could use arbitrary units on a linear scale. However, arbitrary units have no intrinsic meaning and the results of the analysis are, therefore, difficult to interpret and to explain to physicians unfamiliar with decision analysis. Occasionally an ordinal ranking of the outcomes yields dominance of one strategy, that is, all the better outcomes are more likely. However, in many analyses such dominance of one strategy does not exist, in which case the scale has to be proportional to the actual value of the outcomes and an ordinal scale is inadequate. Another approach is to use the survival probability after a fixed period of time. This measure of utility also has associated problems in that, for example, the probability of survival at 5 years does not give information about the survival at 1 year, nor that at 10 years. Furthermore, the patient may value short-term survival more than long-term survival, as is often the case (13).

Life expectancy is a convenient measure of utility, which most people understand intuitively. Life expectancy (LE) summarizes the average future life years a person may expect to live, at a specified age, with or without specified diseases (21). If a group of patients is followed for many years, and the fraction of surviving patients is determined after each year, one can estimate the life expectancy in years of such patients by summing the determined fractions. Life expectancy is usually expressed in units of years.

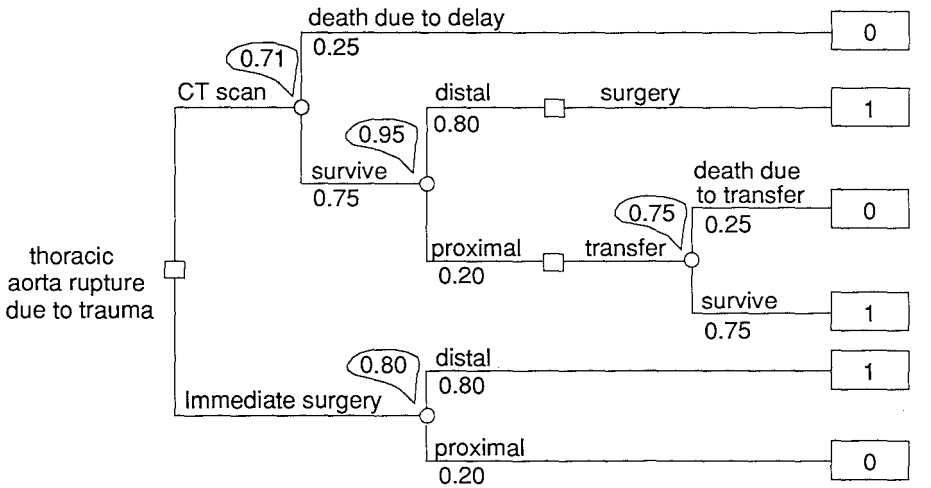


Figure 3. Folding back for the presented example.

Although life expectancy is an objective measure of outcome, patient preferences such as risk aversive and risk seeking attitudes towards life, and the quality of life associated with morbidity, are major factors affecting the patient's perceived value of an outcome (11,12,13,14,16). Most people value the first few coming years to a larger extent than years in the distant future. This attitude is termed risk aversiveness (13). Others, however, are willing to take a fairly large risk now to gain a small increase in life expectancy. This attitude is termed risk seeking. To take into account the patient's attitude towards the quality of life with morbidity, we need to know how he/she values being in an ill state of health (14). The patient's attitude towards ill-health can be quantified by determining the period of time in full health which, for the patient, is equivalent to a year with illness, conventionally expressed in units of quality-adjusted life years (QALY's).

4. Expected utility and folding back

The expected utility of a strategy is the outcome expected on average (21,25). The expected utility is simply the weighted sum of the values of all possible outcomes, with each outcome weighted for the probability that it will occur. The process of calculating the expected utility is called folding back and averaging out (21,25). In the example, folding back for the strategy "immediate surgery" gives an expected utility of 0.80 (= 0.80x1 + 0.20x0) (figure 3). Folding back for the strategy "CT scan" gives an expected utility of 0.71 (= 0.25x0 + 0.75 x (0.80x1 + 0.20 x (0.25x0 + 0.75x1))). The unit of the expected utility of a strategy is the same unit as that in which the outcomes are expressed. In the example, the outcomes are expressed as one if alive and zero if dead and the unit of expected utility is, therefore, the fraction of a cohort expected to be alive.

5. Sensitivity analysis

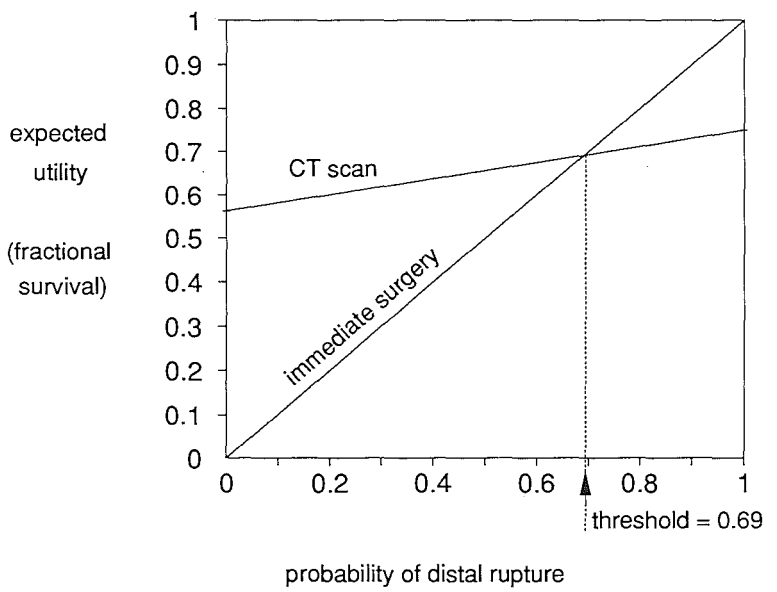


Figure 4. Sensitivity analysis for the probability of a distal rupture of the thoracic aorta.

As noted above, the value of a variable used in an analysis may be subjective, and subject to controversy. Changing the value and recalculating the expected utility may alter the preferred strategy. By repeating the process over a range of values, one performs what is called a sensitivity analysis (21,25). In this example the probability of a distal rupture was estimated at 0.80. If we recalculate the tree for values of this probability from 0 to 1 and graph the expected utility as a function of the probability, we get the graph shown in figure 4. The x-axis represents the analyzed variable, ie. the probability of a distal rupture. The y-axis represents the expected utility in fractional survival. The lines, namely "immediate surgery" and "CT scan", plot the expected utility of the strategy as function of the probability of distal rupture.

If the rupture is likely to be proximal (ie. the probability of a distal rupture is low) performing a CT scan to determine the exact site of the rupture, and transferring the patient if necessary, is better than immediate surgery. If, however, the probability of a distal rupture is high, it is better to proceed immediately to surgery. Patients with a proximal rupture have lower chances of surviving, whatever strategy one chooses, because all patients would die from immediate surgery while transferring the patient brings with it a risk of death. This implies that the expected utility of both strategies increases with increasing probability of a distal rupture.

The point at which the expected utilities are equal for both strategies is called the threshold value for the variable analyzed (21,25). The results are "sensitive" to a variable if a threshold exists, in which case the analyst should try to confirm the exact value of the parameter. The results are "insensitive" to a variable if no threshold exists, in which case the analyst can feel fairly confident of the results. A two-way sensitivity analysis implies calculating the threshold of a variable for various values of a second variable. For example, we could calculate the threshold of the probability of a distal rupture for various values of the probability of death due to delay. This principle can be extended to a three-way, or even a multi-way, sensitivity analysis.

The presented patient underwent a CT scan (figure 5). The images show a distal thoracic aortic rupture. Unfortunately the patient died immediately after the imaging procedure.

Chapter III presents a decision model illustrating the use of a decision tree, in which the choice of workup and treatment for suspected renovascular hypertension is examined. Sensitivity analysis is extensively used to examine the effect of different values of the variables.

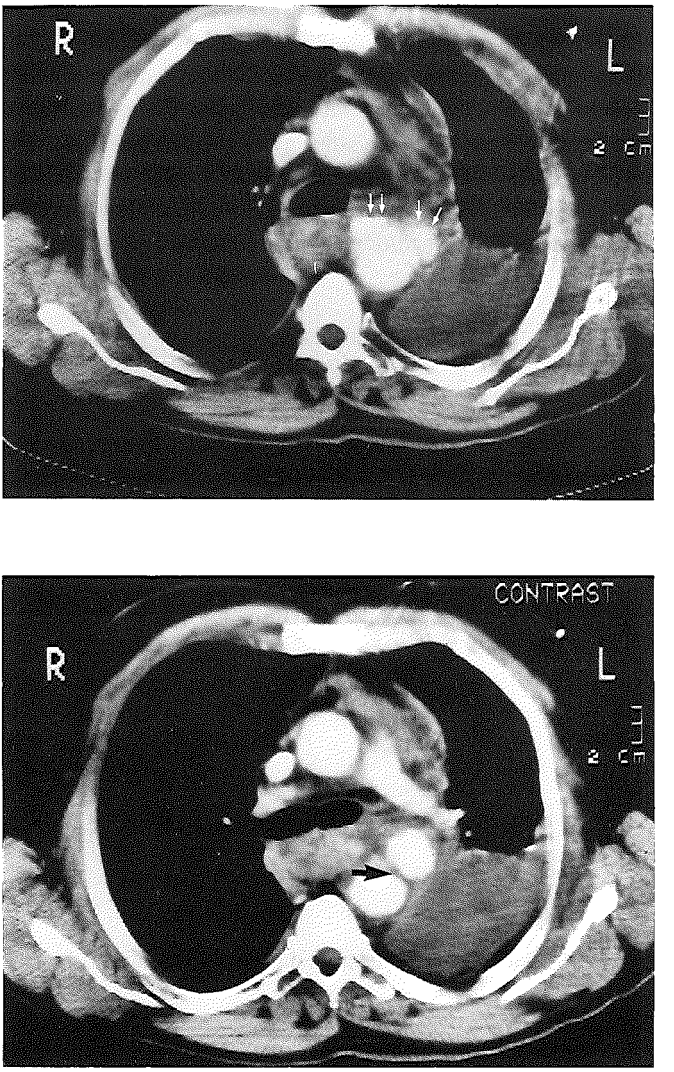


Figure 5. CT scan images of the car accident victim showing distal thoracic aortic rupture. The images demonstrate extravasation of contrast (↓↓) and an intimal flap (➔) distal to the origin of the left subclavian artery.

II. MARKOV MODELS

Markov models are commonly used to model prognosis of chronic diseases. However, problems with a short time horizon may also be conveniently modelled with such models. Markov models are especially useful for modelling prognosis when a risk recurs repetitively over a long period of time, when the likelihood of an event changes over time or when the utility of the associated outcome depends on when the event occurs. In technical terms, a Markov model is convenient when a decision problem is recursive and/or encompasses time-dependent risks.

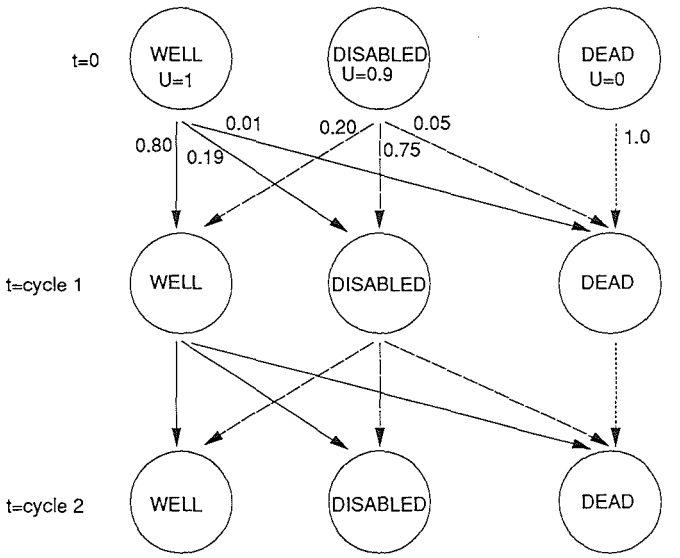


Figure 6. Illustration of a Markov process. U denotes the incremental utility of being in that state for one cycle.

To illustrate the principles of a Markov model, consider a patient with rheumatoid arthritis. Assume we want to calculate the life expectancy of this patient, taking into account the quality of life when disabled by the disease. The patient can be WELL, DISABLED or DEAD, which can be defined as the Markov states of this particular model (figure 6).

Suppose we analyze the prognosis of the patient by increments of 1 year, then the cycle length of the Markov model is 1 year. During each cycle a probability exists that the patient goes from one state to another, called a transition probability. For example (figure 6), if the patient starts out in the WELL state, the (transition) probability that the patient stays in the WELL state equals 0.80. Each state may add a different incremental utility to the overall expected utility. The incremental utility refers to the utility of spending one cycle length in a particular state. For a cycle length of 1 year, the WELL state has an incremental utility of 1 year. The incremental utility of the DEAD state is 0. For the DISABLED state the incremental utility could be a year adjusted for the quality of life for being in the DISABLED state, for example, this could be 0.90 quality-adjusted life years. The DEAD state is called an absorbing state because once the patient is dead, he/she remains in that state and no additional utility is added from being in that state.

Table 1. Calculating quality-adjusted life expectancy with a Markov cohort simulation, assuming the WELL state has an incremental utility of 1 and the DISABLED state has an incremental utility of 0.90

CYCLE	FRACTION in STATE			INCREMENTAL UTILITY		
	WELL	DIS-ABLED	DEAD	WELL	DIS-ABLED	DEAD
0	1	0	0			
1	0.80	0.19	0.01	0.80	0.17	0
2	0.68	0.29	0.03	0.68	0.27	0
3	0.60	0.35	0.05	0.60	0.31	0
4	0.55	0.38	0.07	0.55	0.34	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
40	0.18	0.15	0.67	0.18	0.14	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
UTILITY of STATE				19.83	14.25	0
EXPECTED UTILITY = 19.83 + 14.25 = 34.1 QALY's						

The basic property of a Markov model is that it has no memory for previous states, which is termed the Markovian assumption. In other words, all patients in the DISABLED state have the

same prognosis regardless of their history. Clearly, this assumption is not always valid in clinical problems. However, if the model is not realistic enough and some history needs to be included, one can create an additional state to depict a state of health with a different previous history.

If the transition probabilities are all constant over time, the Markov model is called a Markov chain (or homogeneous Markov model) and the expected utility can be calculated with matrix algebra (3). However, Markov chains are exceptional cases because usually the transition probabilities change over time, in which case the model is called a Markov process. The expected utility of a Markov process may be calculated by a cohort simulation. A cohort simulation begins with an initial distribution of the states. During each cycle the cohort is redistributed according to the transition probabilities. During each cycle the incremental utility of each state, multiplied by the fraction of the cohort in that state, is added to the overall expected utility. For example, if all patients of a cohort start out in the WELL state, the WELL state has an incremental utility of 1 and the DISABLED state has an incremental utility of 0.90, we calculate the quality adjusted life expectancy as in table 1.

It is also possible to model a problem as a combination of a Markov process and a decision tree. Usually the events within one cycle are structured in a decision tree, which is termed the cycle tree. Chapter IV illustrates the application of a Markov process in modelling a problem with a short time horizon and time-dependent probabilities, including a recursive cycle tree in the model.

III. RATES AND PROBABILITIES

In clinical epidemiology there are three basic concepts for expressing a probability. These are:
1) the prior probability and prevalence
2) the cumulative incidence, or cumulative failure function, and
3) the incidence density, hazard function or hazard rate.

All concepts of probability can be derived from these three basic concepts. The mentioned concepts, in my opinion basic to decision analysis, are not found in textbooks on decision analysis. In this section the concepts are described, using the terminology from clinical epidemiology and survival data analysis, and the relevance to decision analysis is discussed.

Note that the term "rate" is generally used in an ambiguous manner. Sometimes the term "rate" refers to an observed relative frequency or proportion (24). In decision analysis the observed relative frequency of events in a group of patients is used to estimate the probability, or cumulative incidence, of an event in the individual. At other times the term "rate" refers to a quantity measured with respect to time (24), referring to the incidence density or hazard rate. Usually the precise meaning of the word "rate" is clear from the context in which it is used. Where necessary, to avoid ambiguity, the terminology described below should be used.

1. The prior probability and prevalence

The prior probability is the probability that a disease D is present at one point in time, which is a unitless entity. The term prevalence is used to mean the probability that a disease D is present at one point in time in a specified population. If the specified population is specific enough for the case at hand, that is the signs and symptoms of the patient are identical to that seen in the specified population, then the prevalence can be used as estimate of the prior probability of disease in the patient. Otherwise, the prevalence must be adjusted to take into account the signs and symptoms of the patient. For example, the prior probability of a renal artery stenosis in a hypertensive female patient of 45 years, will be larger than the prevalence of renal artery stenosis in the general population. (For a discussion on the likelihood ratio and odds see chapter V.)

2. The cumulative incidence or cumulative failure function

The cumulative incidence, or cumulative failure function, is the probability of an event within a set time interval. It is a unitless entity with a value from 0 to 1. The term "failure" means failure to remain in the initial state of health. Expressed as a function of time, the cumulative incidence, or failure function, F(t), is the probability that the event occurs before or at time t (9). If the event is "death" then an estimate of F(t) is

$$\hat{F}(t) = \frac{\text{number of patients dead at } t}{\text{total number of patients}}$$

The value of F(t) is larger than or equal to zero and smaller than or equal to one ($0 \leq F(t) \leq 1$). The cumulative survival function S(t) is the probability that an individual survives longer than t (9). S(t) and F(t) are each others complement:

$$S(t) = 1 - F(t)$$

The survival function depicts what fraction of the initial cohort is alive at a specified time after t = 0. The area under the survival function is the life expectancy of the patient. A life-table is a non-parametric estimate of the survival function, specifying the observed fraction alive after specified time intervals. Parametric survival functions are functions definable by an equation, such as the declining exponential function, the Gompertz function and the Weibull function. Estimates of parametric survival curves have the form of smooth functions, rather than step functions.

"Survival" and "failure" should be considered in a broad sense to allow for functions modelling an event other than death. For example, if we consider the event "spontaneous passage of an ureteral stone", the concept "failure" would mean failure of the stone to remain in the ureter, in other words spontaneous passage. "Survival" would mean that the stone remains in the ureter (see chapter IV).

It is convenient to know the definition of another function, the probability density function, even though this function is used infrequently in clinical decision analysis. Understanding the probability density function helps understanding the hazard function and the associated equations. The probability density function f(t) is the probability of an event in a small interval, expressed per unit time.

From the definition it follows that the probability density function, f(t), is the first derivative of the cumulative failure function, F(t) (9)

$$f(t) = F'(t)$$

f(t) has the following properties:

$$f(t) \geq 0 \text{ for } t \geq 0$$

$$f(t) = 0 \text{ for } t < 0$$

and

$$\int_0^\infty f(t) \cdot dt = 1$$

3. The incidence density, hazard function or hazard rate

The incidence density, or hazard function, is the probability of the occurrence of an event in a small time interval, per unit time, on condition that the event has not occurred before the interval. For example, if the event is death, the condition is that the patient is alive at the beginning of the interval. Other terms used for the same function, if the event is death, are instantaneous mortality rate and force of mortality. The unit in which the incidence density or hazard function is expressed, is events per unit time.

From the definition it follows that the hazard function is the probability density function divided by the survival function.

$$h(t) = \frac{f(t)}{S(t)}$$

As explained before, the probability density function is the first derivative of the cumulative failure function and thus:

$$h(t) = \frac{F'(t)}{S(t)}$$

Because $F(t) = 1 - S(t)$, it follows that $F'(t) = -S'(t)$, and therefore

$$h(t) = \frac{-S'(t)}{S(t)}$$

Integrating the hazard function from 0 to t, we find[1]

$$\int_0^t h(x)dx = -\int_0^t \frac{S'(x)}{S(x)}dx$$

Using the following rule from calculus

$$\frac{d}{dx}\ln x = \frac{1}{x}$$

and assuming that at time t = 0 the whole cohort is in the initial state, that is S(0) = 1, we find

$$\int_0^t h(x)dx = -\ln S(t)$$

From this we derive

$$S(t) = e^{-\int_0^t h(x)dx}$$

and

$$F(t) = 1 - e^{-\int_0^t h(x)dx}$$

[1]Note that h(t) and h(x) are the same function. x is here used to denote the function variable to distinguish it from the integration variable t.

If the incidence density or hazard function is a constant, $h(x) = \mu$, then

$$S(t) = e^{-\mu t}$$

$$F(t) = 1 - e^{-\mu t}$$

IV. MODELLING TIME-DEPENDENT TRANSITION PROBABILITIES IN MARKOV PROCESSES

In a simple decision tree, time-dependent probabilities are conveniently modelled as the cumulative incidence or cumulative failure function. The probability that the event has occurred before or at time t is expressed as F(t). Modelling time-dependent transition probabilities in Markov processes necessitates some extra thought. From a literature review and other gathered data, it is usually practical to derive an incidence density or hazard function of the event. The question is how to apply this function in the Markov process. Although the translation of a constant hazard rate to a transition probability has been described before (3), the translation of a non-constant hazard function to a time-dependent transition probability function has not been described in detail in the literature. The relevant equation is derived and, for a constant hazard rate, shown to be equivalent to the equation previously described in the literature.

Assume you have two Markov states, survival or death. The transition probability $P_{S \to S}$ from the state survival at time t_i to the same state (survival) in the next cycle, at time t_{i+1} is

$$P_{S \to S} = \frac{S(t_{i+1})}{S(t_i)}$$

$$= \frac{e^{-\int_0^{t_{i+1}} h(x)dx}}{e^{-\int_0^{t_i} h(x)dx}}$$

$$= e^{-\int_{t_i}^{t_{i+1}} h(x)dx}$$

Because F(t) = 1 - S(t), the transition probability $P_{S \to F}$ from the state survival at time t_i to the dead state (ie. failure) at time t_{i+1} is

$$P_{S \to F} = 1 - e^{-\int_{t_i}^{t_{i+1}} h(x)dx}$$

The derived equation translates the incidence density or hazard function to a cumulative incidence or failure function for the time interval t_i to t_{i+1}, and is the probability that should be used in the Markov process. If the hazard function is a constant μ and the cycle length is Δt, then the transition probability is

$$P_{S \to F} = 1 - e^{-\int_{t_i}^{t_{i+1}} \mu dx}$$
$$= 1 - e^{-\mu \cdot \Delta t}$$

Often the hazard function is not constant over time, but if the Markov cycle length Δt is small, we can estimate the hazard function with a stepwise function. That is, during the Markov cycle from t_i to t_{i+1} we estimate the hazard function with a constant equal to $\mu\left(t_i + \frac{1}{2}\Delta t\right)$. As long as $\mu\left(t_i + \frac{1}{2}\Delta t\right) \cdot \Delta t$ is small, this is an adequate estimate. The precise criterion you use to determine if the estimate is adequate, depends on how accurate the calculations have to be. A convenient way to implement this, is to program the tree so that the cycle length is automatically shortened if $\mu\left(t_i + \frac{1}{2}\Delta t\right) \cdot \Delta t$ becomes bigger than the criterion value.

V. MODELLING LIFE EXPECTANCY

1. The declining exponential approximation of life expectancy: DEALE-method

The DEALE-method, introduced by Beck et. al., assumes that the incidence density, hazard function or hazard rate for the event death is a constant:

$$h(t) = \mu$$

μ is also called the rate of failure, mortality rate or force of mortality and is expressed in events per unit time.

The corresponding survival and cumulative incidence or failure functions are:

$$S(t) = e^{-\mu t}$$

$$F(t) = 1 - e^{-\mu t}$$

The convenience of the DEALE-method lies in the fact that life expectancy equals the reciprocal of the patient's hazard rate, ie. if the patient's hazard rate is a constant μ then:

$$LE = \int_0^\infty S(t)dt$$

$$= \int_0^\infty e^{-\mu t}dt$$

$$= \left[-\frac{1}{\mu}e^{-\mu t}\right]_0^\infty$$

$$= \frac{1}{\mu}$$

μ is the patient specific hazard or mortality rate. If the patient has no particular disease, μ is the average mortality rate for the rest of the patient's life, which is the reciprocal of the life expectancy, specific for age, sex and race. This average mortality is also called the age, sex and race specific mortality rate. If the patient has a disease, we add the excess mortality rate specific for the disease to the age, sex and race specific mortality rate to calculate the patient specific mortality rate. Taking the reciprocal of the patient specific mortality rate we derive the patient specific life expectancy. In equations:

LE_{ASR}	life expectancy specified for age, sex and race, found in life table statistics
$\mu_{ASR} = \frac{1}{LE_{ASR}}$	average mortality rate, age, sex and race specific
μ_{DS}	excess mortality rate, disease specific
$\mu = \mu_{ASR} + \mu_{DS}$	summation gives patient specific mortality rate
$LE = \frac{1}{\mu}$	reciprocal gives patient specific life expectancy

The declining exponential function assumes that the age of the patient does not affect the hazard rate, ie. the hazard rate is constant for the rest of the patient's life. This assumption grossly simplifies reality. Distinguishing the average mortality rate calculated with the DEALE-method from the instantaneous mortality, as given in life tables, helps understand the effect of the assumption. The instantaneous mortality rate increases with age, implying that for young people the average mortality rate calculated with the DEALE overestimates the instantaneous mortality rate, and for old people the average mortality rate underestimates the instantaneous mortality rate. However, in spite of the gross assumption made, the DEALE-method is a convenient estimation method.

2. Disease specific excess mortality rates

Mortality rates for a specific disease are often reported as overall mortality rates as observed in a study population, including mortality from other causes, rather than as excess mortality rates for the disease only. The reported rates must, therefore, be adjusted for the mortality rate of a healthy population similar to the study population with respect to age, sex and race (1,2). Two methods can be applied, which, as will be shown, are equivalent.

a) Correcting for mortality due to other causes by subtraction

Of a study population with disease D the surviving fraction at time t is $S_D(t)$. The incidence density or instantaneous mortality rate of this population is

$$\mu_D = -\frac{d}{dt}\ln S_D(t)$$

To correct for mortality due to other causes we subtract the mortality rate of a population similar to the study population with respect to age, sex and race, but healthy, and followed for the same length of time as the study population:

$$\mu_{DS} = \mu_D - \mu_H$$

μ_H is derived from survival data in life tables. If the average age of the study population at the beginning of the study is AGE and patients are followed for a period Δt, the average age at the end of the study will be $AGE + \Delta t$. The mortality rate for a similar but healthy population μ_H is the cumulative survival at $t = AGE + \Delta t$ divided by the cumulative survival at $t = AGE$ (as found in life tables). Expressed in an equation:

$$\mu_H = \frac{S_H(AGE + \Delta t)}{S_H(AGE)}$$

where $S_H(t)$ is the survival function of the similar healthy population. The survival probability derived from life tables is based on the general population and will, therefore, include among the so called "similar healthy population", a certain fraction who have the particular disease we are interested in. Strictly speaking, we should exclude from the "similar healthy population" patients with the disease we are interested. However, this is cumbersome and usually not done in practice.

Beck et.al. describe the above method in their original article (2). In their example they derive μ_H by taking the reciprocal of the life expectancy of a healthy population, similar in age and sex to the study population. However, this overestimates the actual μ_H, and therefore underestimates the corresponding excess mortality rate. It is, therefore, prudent to derive μ_H from life tables, instead of taking the reciprocal of life expectancy.

b) Correcting for mortality due to other causes by using relative survival

Survival is sometimes presented as relative survival, that is, relative to a similar but healthy population. This means that overall survival of the study cohort is divided by survival of a population similar in age, sex and race, but healthy, and followed for the same length of time. If $S_D(t)$ and $S_H(t)$ are the survival functions of the study population with the disease and a similar healthy population respectively, then relative survival RS is

$$RS(t) = \frac{S_D(t)}{S_H(t)}$$

From the relative survival we calculate the disease specific excess mortality rate μ_{DS} with the equation

$$\boxed{\mu_{DS} = -\frac{d}{dt}\ln RS(t)}$$

The method described under b) is, in fact, equivalent to the method under a), because

$$\begin{aligned}\mu_{DS} &= -\frac{d}{dt}\ln RS(t) \\ &= -\frac{d}{dt}\ln\left(\frac{S_D(t)}{S_H(t)}\right)\end{aligned}$$

$$
\begin{aligned}
&= -\frac{d}{dt}(\ln S_D(t) - \ln S_H(t)) \\
&= \left(-\frac{d}{dt}\ln S_D(t)\right) - \left(-\frac{d}{dt}\ln S_H(t)\right) \\
&= \mu_D - \mu_H
\end{aligned}
$$

It is convenient to understand both methods, and their equivalence, because data is sometimes presented as overall survival and sometimes as relative survival.

3. The Gompertz model

A more general model than the DEALE-method is the Gompertz model, in which the hazard rate is not a constant but instead an increasing exponential function of time (or age) (9):

$$h(t) = e^{\lambda + \gamma t}$$

The corresponding cumulative incidence or failure function is:

$$F(t) = 1 - e^{-\frac{e^{\lambda}}{\gamma}(e^{\gamma t} - 1)}$$

For $\gamma = 0$, h(t) reduces to a constant, e^{λ}, and the Gompertz is equivalent to the DEALE.

4. The Weibull model

Another generalization of the declining exponential (DEALE) method is the Weibull model, in which the hazard rate is not a constant, but instead a function of time (or age) (9):

$$h(t) = \lambda\gamma(\lambda t)^{\gamma - 1}$$

The corresponding cumulative incidence or failure function is:

$$F(t) = 1 - e^{-(\lambda t)^{\gamma}}$$

For $\gamma = 1$, h(t) reduces to a constant λ and the Weibull is equivalent to the DEALE.

VI. MODELLING PATIENT PREFERENCES

Risk aversion versus risk seeking attitudes, and the quality of life with morbidity, are patient preferences which may be taken into account when analyzing a decision problem.

1. Quality of life with morbidity

In assessing patient preferences with respect to quality it is important to distinguish between the anticipated quality of life with morbidity of a hypothetical event as perceived by someone faced with a choice, and the average experienced quality of life of a group of people with morbidity. An analysis for an individual patient incorporates the individual anticipated quality of life, in other words the preference of the patient. Generic models and cost-effectiveness analyses incorporate the average quality of life experienced by patients with morbidity.

The quality factor is the quantity that identifies how much a patient values a year with morbidity compared to a year in full health. This factor is multiplied by the life expectancy to calculate the quality-adjusted life expectancy:

$$LE \times quality\ factor = quality\ adjusted\ LE$$

which is expressed in units of quality-adjusted life years. However, the unit "quality-adjusted life years" is confusing because it does not distinguish between the individual anticipated quality of life and the average experienced quality of life. We propose a new unit, "preference-adjusted life years" (PALY's), to express the life expectancy adjusted for the individual anticipated quality of life.

To determine the anticipated quality of life three basic methods are in use: the direct scaling (or category) method, the time trade-off method and the standard reference gamble. With the direct scaling method one asks the patient to mark his/her perceived value of the states of ill-health on a linear scale from the worst outcome, usually death, to the best outcome, usually full health, placing the other states in between (25) proportional to their value. Although easy to conceptualize, the direct scaling method is inaccurate (18). Towards the outer limits of the scale the outcomes tend to be valued further apart than using the other two methods. For quality factors in the upper part of the scale, the direct scaling method results in lower quality factors whereas for quality factors in the lower part of the scale, the direct scaling method results in higher quality factors than using the other two methods.

The time trade-off method directly assesses what length of time in full health is equivalent to a period of ill-health. The question to the patient is "what is the shortest period of lifetime you would accept in exchange for your life expectancy now, with ill-health?". For example, if a patient in ill-health has a life expectancy of 10 years, but is willing to accept 4 years of full

health in exchange, then 1 year of ill-health is equivalent to 0.4 quality-adjusted life years (21,25). In practice it is easiest to apply this information by means of a "quality factor" (a measure for the quality of life), which in this case is 0.4, and subsequently multiply life expectancy with the quality factor. The time trade-off method is fairly easy to understand for most patients, however, the results are usually higher than what one would obtain using the direct scaling method and lower than using the standard reference gamble (18).

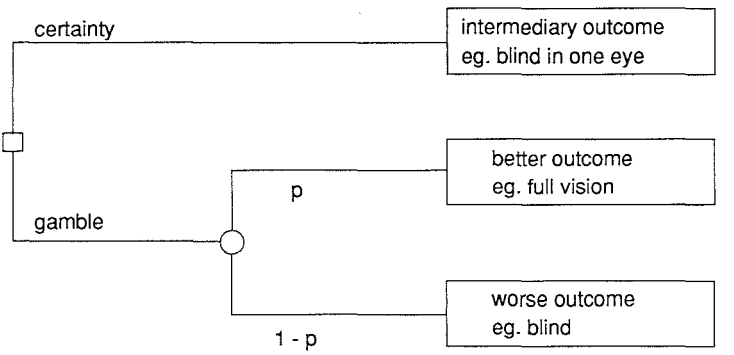


Figure 7. The standard reference gamble.

The standard reference gamble is considered the "gold standard" when assessing quality of life. (The method is also used to determine risk aversive and risk seeking attitudes, see section VI.2.) The question posed to the patient is to choose between 1) a certain intermediary state, for example, blind in one eye or 2) a gamble between a better outcome, for example full vision, with probability p and a worse outcome, for example blind, with probability 1-p (21,25) (figure 7). If the patient is indifferent to the choice for a probability p of say 0.4, then the quality factor of the intermediary outcome compared to the better and worse outcome, is 0.4.

It remains a difficult task to obtain a quality factor for a particular patient because of the difficulty of the questions posed, the confrontation with the possible consequences of the disease and because the elicitation of preferences takes a lot of time.

When constructing generic models or performing cost-effectiveness analysis as opposed to analyzing individual decisions, it is not realistic to use one patient's preferences as representative for a whole group. In this case the average experience of a group of similar patients is used. This may be done by questioning patients with the state of ill-health and scoring their experience. Rosser and Kind have developed a rating scale for functional impairment and distress and have determined the corresponding quality factors (19). They have classified states of illness according to disability and distress. Disability and distress are scored as follows:

DISABILITY

1. no disability
2. slight social disability
3. sever social disability and/or slight impairment of work
4. choice of work severely limited
5. unable to undertake paid employment or continue any education, confined to home except for short outings
6. confined to (wheel) chair
7. confined to bed
8. unconscious

DISTRESS

1. no distress
2. mild
3. moderate
4. severe

Table 2. Table of ratio scale for different levels and combinations of disability and distress (modified from reference 19).

DISABILITY	DISTRESS			
	1	2	3	4
1	1.000	0.995	0.990	0.967
2	0.990	0.987	0.973	0.933
3	0.980	0.972	0.956	0.913
4	0.964	0.957	0.942	0.870
5	0.946	0.935	0.900	0.700
6	0.875	0.845	0.680	0
7	0.678	0.564	0	-1.486
8	-1.029			

30 states of health representing different levels and combinations of disability and distress were presented to 70 subjects, including patients, doctors, nurses and healthy volunteers. Subjects were asked to rank the states from least ill to more ill, rank 1 representing the least ill state. Subsequently they were asked to assign values to represent how much more ill someone is, for example, in the state ranked 2 compared to the state ranked 1, assuming the states are permanent. The given value had to represent their judgment of, for example, the proportion of resources

they felt should be allocated to patients in the state ranked 2, compared to patients in the state ranked 1. The subjects were interviewed at length and asked to re-evaluate their judgments until they felt comfortable with the given results. They were asked to include "well" and "dead" in the ranking order and assign values to these in accordance with their other assigned values. The values given by these subjects can be used to calculate a ratio scale for various states of health as defined by the disability and distress scores. The derived ratio scale can be used as quality factors for the various states of health, as determined by disability and distress (table 2). Distress can be regarded in a broad sense including pain. A limitation of this scale is the fairly wide range of values given, which depended on whether the subject was a patient, doctor, nurse or healthy volunteer. However, this apparent limitation can be used to the decision analyst's advantage by performing a sensitivity analysis for the range of values, or choosing the results from the group most representative for the analysis.

2. Risk aversion and risk seeking attitudes

A year in the distant future may be perceived by the patient as having less value than a year now. For example the choice between surgical treatment or radiation for lung cancer is often based on the 5 year survival, 5 year survival of surgery being greater than the 5 year survival of radiation. However, patients might be averse to the immediate risk of surgical mortality and consider life during the next few months more important than later years (12,13,16). Such a patient would opt for certain short survival in favour of a gamble between death now and longer survival. This attitude to life is called risk aversion.

To deal with this phenomenon in decision analysis we construct a utility curve, that is a curve of the value of varying periods of survival. The curve is constructed by presenting hypothetical choices between a certain short survival and a gamble on longer survival, called the standard reference gamble.

For example, consider the following hypothetical choice:

> Looking at a glass ball we predict that you will live 3.5 years. There exists a magic pill that can make you live 7 years. However, this pill has one serious side effect: 50% of people who take it die within 3 days. Do you want to take the magic pill ? The choice is therefore:
>
> A) no risk now and living 3.5 years for certain
>
> B) a very risky treatment with a chance of dying now of 0.50, but if the treatment is successful living 7 years

Many people would choose A even though the life expectancy of the two options is equal. If A was a certain survival of, say, 2 years many would still choose A. This is called risk aversion, that is one would rather have the certainty of living the shorter period of 2 years than taking the gamble between a risk now of 50% and a longer life expectancy of 7 years. The breakeven point, that is the value of the certain period of survival of A for which A and B are perceived as equal,

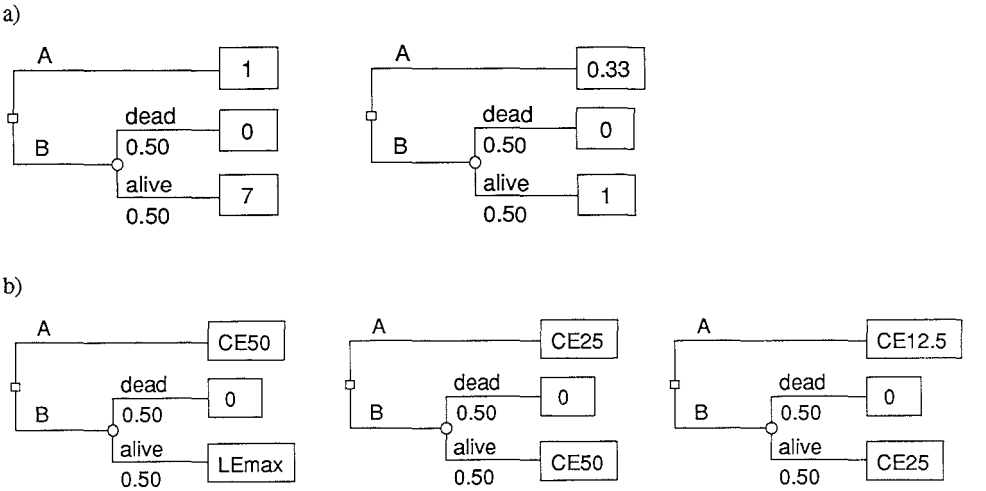


Figure 8. Schematic representation of the procedure to derive a subject's certainty equivalents. a) shows the choices for the example presented and b) shows the choices using conventional notation. CE50 = certainty equivalent 50, CE25 = certainty equivalent 25, CE12.5 = certainty equivalent 12.5, LEmax = maximum life expectancy.

is called the certainty equivalent. More specifically, the derived value is the certainty equivalent 50 (CE50), that is the period of certain survival equivalent to 50% of the uncertain longer survival. Say CE50 is 1 year, then the next task would be to estimate x, posing the hypothetical choice:

A) no risk now and living x years for certain

B) a very risky treatment with a chance of dying now of 0.50, but if the treatment is successful living 1 year

The derived value of x is called the certainty equivalent 25 (CE25). We can proceed further in a similar fashion to derive the complete utility curve. The procedure is presented schematically in figure 8.

An example of a completed utility curve is given in figure 9, whereby the values between the known certainty equivalents are derived by linear interpolation. Not everybody has a risk aversive attitude to life: some people are risk seeking, preferring a 50-50 gamble between death now and 7 years survival to certain survival of 3.5 years. The utility curve of a risk seeking person lies below the diagonal. Note that someone who is risk neutral would have a straight utility curve along the diagonal.

A practical approach to take risk aversiveness into account in a decision model is to assume that the utility curve has a parametric form (16), for example the form:

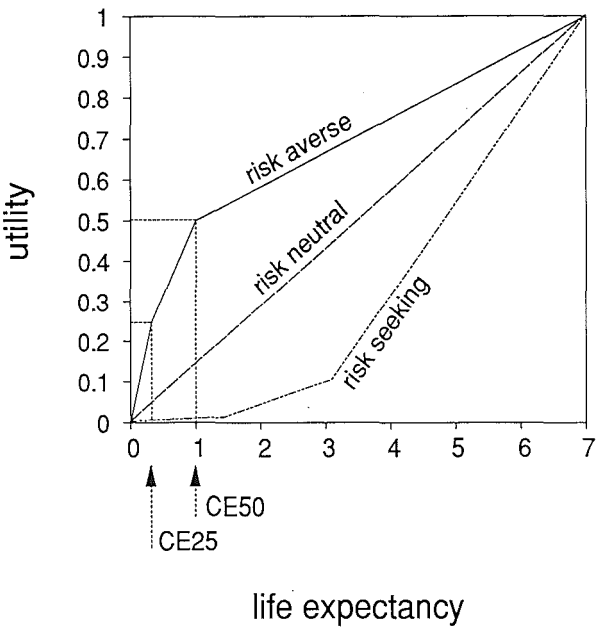


Figure 9. Utility curve for a risk averse person with LEmax=7 years, CE50=1 year and CE25=0.33, with linear interpolation between the known certainty equivalents. Sample curves for risk seeking and risk neutral subjects are also shown.

$$U(t) = k \cdot (1 - e^{-r \cdot t})$$

where U(t) is the utility of the period of survival t, k is a scaling constant and r is the aversion rate. To derive the values of k and r we need to know two (non-zero) utilities: for example, if we know the period of survival with utility 1 (that is the maximum life expectancy LEmax) and the period of survival with utility 0.5 (that is the certainty equivalent CE50), we can calculate the parameters k and r and the utility curve will be defined. k and r are calculated by substituting the two known utilities in the formula defining the curve, from which we get two equations with two unknowns, which can be solved. For example, if LEmax=7 years and CE50=1 year, then the equations are:

$$U(7) = 1 = k \cdot (1 - e^{-7 \cdot r})$$

$$U(1) = 0.5 = k \cdot (1 - e^{-1 \cdot r})$$

Solving for k and r we get k=1.0082 and r= 0.685. The utility curve U(t) defined by these values for k and r is given in figure 10.

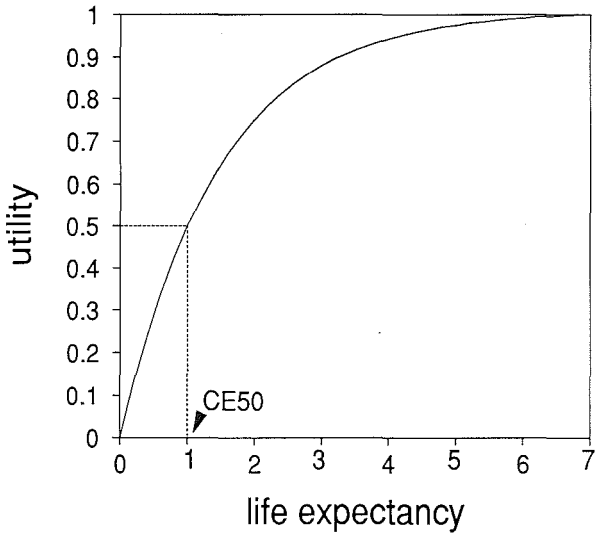


Figure 10. Utility curve for a risk averse person using the parametric form $U(t) = k \cdot (1 - e^{-r \cdot t})$ with k=1.0082 and r=0.685, derived by substituting LEmax=7 years and CE50=1 year.

Another possible attitude towards life, not described previously in the literature, is one that varies between risk seeking and risk aversion depending on the period of survival. One could argue that a very short period of survival, say 1.5 years, is too short to do anything worth-while and that a gamble between a 50% chance of dying now and living 3 years would be preferred. The same person could however be risk averse when asked to choose between a certain survival of 3.5 years and the 50-50 gamble between dying now and living 7 years. The resulting utility curve would be below the diagonal for short periods of survival and above the diagonal for longer periods of survival. For example, an intern would probably be risk seeking for short life expectancy but risk averse for longer periods of life expectancy. Young people are more likely to have a variable attitude towards risk than older people. The curve could be modelled with a smoothed sigmoid curve, for example a cumulative normal probability function of the general form:

$$U(t) = \Phi\left(\frac{t - CE50}{s}\right)$$

where U(t) is the utility of the period of survival t, CE50 is the certainty equivalent 50 and s is a scaling constant. To derive the general form of this function we would need to know the CE50 and the certainty equivalent of a utility other than 1 or 0 (note: U(LEmax)=1 or U(0)=0). For example, if somebody has a CE50 of 3 years and a CE25 of 2.5 years, then substituting these values and solving for s gives s=0.746. The corresponding curve is given in figure 11.

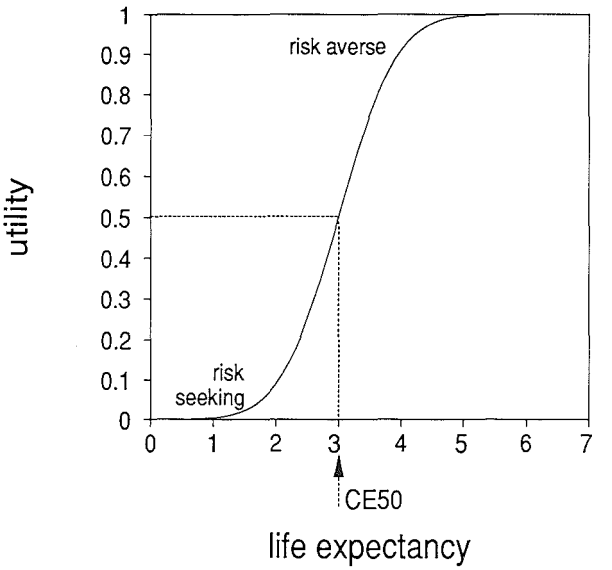


Figure 11. The "intern's utility curve": utility curve for varying attitudes depending on the period of survival. This person is risk seeking for short periods of survival but risk averse for longer periods.

One of the problems with deriving someone's utility curve is the framing of the questions as this influences risk aversive versus risk seeking attitudes (11,23). The usual pattern is one of risk aversion in choices concerning gains and risk seeking in choices concerning losses. Furthermore "losses loom larger than gains" (quote 23), that is, a year of life lost has more impact than a year of life gained. If the decision is the same irrelevant of how the choices are framed, one can be confident the answer reflects the true preference. However, in eliciting patient's preferences one will often find conflicting answers and this issue is as yet unresolved.

The newly introduced unit preference-adjusted life years (PALY's) is also useful when including attitudes towards risk. A utility curve converts periods of survival to a utility on a scale from zero to one, adjusted for risk aversive and risk seeking attitudes. Multiplying the calculated

utilities by the maximum life expectancy will yield life expectancy adjusted for the patient's attitude towards risk and, thus, if life expectancy has also been adjusted for anticipated quality of life, may be expressed in units of preference-adjusted life years.

VII. THE AXIOMS OF EXPECTED UTILITY THEORY AND AN ALTERNATIVE THEORY: PROSPECT THEORY

Expected utility theory is the basis of decision analysis. In principle expected utility theory provides a "prescriptive" approach to decision making under uncertainty. Whether the "prescribed" strategy is followed depends not only on the results of the analysis, but also on the confidence the clinician has in the analysis, and possibly factors not included in the analysis, such as patient preferences or financial costs. Observations of preferences under uncertainty, however, have shown that expected utility theory is not an adequate "descriptive" model of decision making (8,23). Various underlying assumptions of expected utility theory are violated by decision makers when faced with choices under uncertainty.

The underlying assumptions, or axioms, of expected utility theory are (8):

1) the expectation principle: the overall utility of a decision equals the expected utility of its outcomes, that is the summation of the utility of the outcomes each weighted for the probability that the outcome will occur.

2) asset integration: utility is a value of the final state, rather than a relative gain or loss compared to one's current position

In economics, there is a third assumption:

3) risk aversiveness: a subject is risk averse if he prefers a certain outcome to any probable outcome with the same expected utility.

In medical decision making risk aversiveness is usually the observed preference of the patient and is, therefore, often taken into account when constructing the decision model. However, in medical decision models risk aversiveness is not considered an axiom.

In studying subjects' preferences under uncertainty, three effects have been observed which violate the basic principles of expected utility theory, the reflection effect, the certainty effect and the isolation effect (8,11,23). Rational decision making would imply that choices do not depend on the framing of the questions. However, presenting options in terms of losses instead of gains, for example, in terms of mortality rates instead of survival rates, reverses the preference for the majority of tested subjects (11,23). In general, people tend to be risk averse when options involve gains, and risk seeking when options involve losses. This is termed the reflection effect

(8): that is, the reflection of options around zero reverses the preference. Furthermore, a difference is valued greater when it is framed as a loss than when it is framed as a gain, in short "losses loom larger than gains" (23).

Another violation of expected utility theory is the certainty effect. According to expected utility theory if X is preferred to Y, then pX is preferred to pY, for any probability p. However, if X is a certain gain and Y is an uncertain gain, then reducing the probabilities of both outcomes to say, one quarter of the original probabilities, reverses the preference for the majority of subjects. For example, given the options:

X: \$ 3000 for certain Y: 0.80 chance of winning \$ 4000

the majority of questioned subjects chose X, whereas given the following options:

pX: 0.25 chance of winning \$ 3000 pY: 0.20 chance of winning \$ 4000

the majority of subjects preferred pY (8). This observation can, however, not be explained by the notion that certainty in itself is desired: instead certainty appears to increase the utility of gains, but also increases the disutility of losses. Furthermore, when the probability of a gain is large, subjects tend to choose the option where a gain is more probable, while if the probabilities are very small, they choose the option that offers the larger gain (8).

A third observation is the isolation effect: if options have common components these are disregarded and decisions are made on the basis of the components that are different. Apparently common components are considered as part of the current asset position and decisions are based on the changes from the current position rather than the final states (8).

To account for the observations made of preferences under uncertainty, an alternative to expected utility theory has been proposed, namely prospect theory (8). Prospect theory attempts to model decision making as observed by empirical data. Each probability is multiplied by a decision weight, which is a measure of the impact of the probability on the value of the option. The decision weights for each probability together define the "weighting function". Furthermore, each outcome is given a value (described by the "value function") depending on the subjective utility of the outcome. The value function is defined by deviations from a reference point.

Prospect theory could be considered a generalization of expected utility theory in that the expectation principle is not an axiom, and instead a weighting function is used for the probabilities involved. Another distinguishing feature is that it focuses on relative differences from an asset position instead of considering final states.

Important to keep in mind is that prospect theory is a descriptive model. This is essentially different from decision making based on expected utility theory, which is in principle prescriptive, as opposed to descriptive. The difference is essential because we have no way of determining which theory is the "correct" one, that is, which theory will provide optimal decision

making. Although prospect theory describes the human decision making process more adequately, we are not at all certain that human beings make optimal decisions. To the contrary, the empirical data (22) shows that our estimates of probabilities are inaccurate. If our estimated probabilities are incorrect, we are likely to be incorrect in interpreting the impact of probabilities. In addition, overwhelming amounts of information cannot be combined by the human mind (5,15). Furthermore, the described reflection effect can be interpreted in two ways: it is a violation of expected utility theory, therefore questioning the axiom's of the theory, but it also describes the irrationality of our decisions, in that decisions depend on the framing of the question.

Summarizing, the observed violations of expected utility theory may be used as an argument either for or against expected utility theory, depending on one's inclination and point of view. Inspite of the described deficiencies and limitations of decision theory, decision analysis helps with the task of decision making under uncertainty by making the process explicit. It assists in structuring the overwhelming amount of medical information that exists, gives insight into the problem and identifies the trade-offs involved and information needed. "Decisions must and will be made. If they are not made actively, they will be made by default" (4), and, thus, it seems worthwhile to explore and utilize every means that will help us making those decisions.

VIII. REFERENCES

1. Beck JR, Kassirer JP, Pauker SG. A convenient approximation of life expectancy (the "DEALE"). I. Validation of the method. Am J Med 1982; 73: 883-888.

2. Beck JR, Pauker SG, Gotlieb JE, Klein K, Kassirer JP. A convenient approximation of life expectancy (the "DEALE"). II. Use in Medical Decision-Making. Am J Med 1982; 73: 889-897.

3. Beck JR, Pauker SG. The Markov process in medical prognosis. Med Decis Making 1983; 3: 419-458.

4. Dawson NV, Arkes HR. Systematic errors in medical decision making: judgment limitations. Journal of General Internal Medicine 1987; 2: 183-187.

5. de Dombal FT. Picking the best tests in acute abdominal pain. Journal of Royal College of Physicians of London, 1979; 13/4: 203-208.

6. Grainger RG, Allinson DJ. Diagnostic Radiology. 1 ed. Churchill Livingstone, Edinburgh 1986.

7. Henrik R Wulff. Rational Diagnosis and Treatment: an introduction to clinical decision making. 2 ed, 1981. Blackwell Scientific Publications, Oxford.

8. Kahneman D and Tversky A. Prospect theory: an analysis of decision under risk. Econometrica 1979; 47/2: 263-291.

9. Lee ET. Statistical Methods for Survival Data Analysis. Lifetime Learning Publications, Wadsworth Inc., Belmont, California, 1980.

10. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975; 293: 211-215.

11. McNeil BJ, Pauker SG, Sox HC, Tversky A. On the elicitation of preferences for alternative therapies. N Engl J Med 1982; 306: 1259-1262.

12. McNeil BJ, Pauker SG. The Patient's Role in Assessing the Value of Diagnostic Tests. Radiology 1979; 132: 605-610.

13. McNeil BJ, Weichselbaum R, Pauker SG. Fallacy of the five-year survival in lung cancer. N Engl J Med 1978; 299; 1397-1401.

14. McNeil BJ, Weichselbaum R, Pauker SG. Speech and Survival: Tradeoffs between quality and quantity of life in laryngeal cancer. N Engl J Med 1981; 305; 982-987.

15. Miller GA. The magical number seven, plus or minus two. Psychol Rev 1956; 63: 81-97.

16. Pauker SG, McNeil BJ. Impact of patient preferences on the selection of theapy. J Chronic Dis 1981; 34: 77-86.

17. Raiffa H. Decision analysis: Introductory lectures on choices under uncertainty. 1ed 1968, Random House, New York.

18. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for Health Outcomes: comparison of assessment methods. Med Decis Making 1984; 4: 315-329.

19. Rosser R, Kind P. A scale of valuations of states of illness: is there a social concensus? Int J Epidemiol 1978; 7: 347-358.

20. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Little, Brown and Company.

21. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Butterworth Publishers, Boston, 1988.

22. Tversky A and Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science 1974; 185: 1124-1131.

23. Tversky A and Kahneman D. The Framing of Decisions and the Psychology of Choice. Science 1981; 211: 453-458.

24. Webster's II. New Riverside Dictionary. Berkley Books, New York, 1984.

25. Weinstein MC and Fineberg HV. Clinical Decision Analysis. 1980. Saunders WB Co. Philadelphia.

Chapter III

MEDICAL DECISION ANALYSIS: RENOVASCULAR HYPERTENSION[1]

I. INTRODUCTION
II. THE STRUCTURE OF THE PROBLEM
III. SUMMARY OF AVAILABLE DATA AND ASSUMED PROBABILITIES
1. The prior probability of renal artery stenosis and renovascular hypertension
2. Sensitivity and specificity of diagnostic tests
3. Complications of diagnostic tests
4. Surgery
5. Percutaneous transluminal angioplasty (PTA)
IV. ASSIGNMENT OF UTILITIES
1. The DEALE-method
2. The excess mortality rates
V. RESULTS
1. Calculating the expected utilities of the strategies
2. Sensitivity analysis
VI. COMMENT
VII. REFERENCES

ABSTRACT

A decision analysis is presented for the choice of diagnostic workup and therapeutic intervention in a patient with hypertension possibly caused by renal artery stenosis. The outcome values are expressed as life expectancies. The strategy chosen is the one that maximizes life expectancy.

---

[1]Co-author: J Lubsen, Center for Clinical Decision Analysis, Erasmus University and Dijkzigt Hospital, Rotterdam.
Retyped version of publication in Journal of Medical Imaging 1988; 2: 61-70. See also Chapter IX paragraph I.6.

I. INTRODUCTION

Decision analysis helps to structure clinical problems, to integrate available information and to select the optimal strategy. Hypertension, possibly caused by renal artery stenosis, presents decision problems as far as the selection of diagnostic tests and appropriate clinical management are concerned. Radiologists are often consulted about the controversial issue of which diagnostic tests to perform and the feasibility of angioplasty (1).

In this paper the problem is analyzed using a decision theoretical approach (2). The necessary data were derived from the literature. The outcome values of the various decisions are expressed in life expectancy. The model is applied to a particular patient.

CASE PRESENTATION

A 55 year old woman is seen with severe hypertension. The blood pressure measures 245/140 mmHg (average of the first two measurements, phase V diastolic). The physical examination gives no information as to the possible cause. No biochemical abnormalities are found. Treatment is started with sodium intake restriction, a beta-blocker and a diuretic. The average blood pressure drops to 200/100 mmHg. A radiologist is consulted.

II. THE STRUCTURE OF THE PROBLEM

The following questions have to be answered:

1. Would intravenous urography (IVU), renography (RG), intravenous digital subtraction angiography (DSA), arterial angiography (AG), or a combination of these be the best diagnostic workup?
2. If renovascular disease is subsequently diagnosed, does percutaneous transluminal angioplasty (PTA) offer a better chance of cure than surgical treatment?

In figure 1 the clinical strategies with their corresponding possible outcomes are summarized in the form of a decision tree. Diagnostic tests considered are intravenous urography, renography, intravenous digital subtraction angiography, arterial angiography and reasonable combinations of these tests. At each stage the workup is continued if the previous test result is positive for renal artery stenosis. If the test result is negative, the hypertension is considered idiopathic and medication is the treatment choice. If the workup results in the diagnosis renal artery stenosis percutaneous transluminal angioplasty or surgical intervention are the therapeutic options considered.

III. SUMMARY OF AVAILABLE DATA AND ASSUMED PROBABILITIES

A summary of the available data used in this analysis is given in Table I.

1. The prior probability of renal artery stenosis and renovascular hypertension:

The prior probability of renal artery stenosis depends on the population the patient originates from. The prevalence of renal artery stenosis ranges from 0.2% of unselected hypertensive patients (diastolic blood pressure of 95 mmHg) registered in a large general hospital to about 4.5% in a large referral clinic (1). Higher prevalences have been found by using stricter selection criteria (1). For the case presented a prior probability of the presence of renal artery stenosis is assumed to be 0.03 (=3%).

Between 60 and 70 % of renal artery stenoses detected are atherosclerotic, the remainder are due to fibromuscular dysplasia. Age at onset of hypertension and sex determine the probability of a stenosis being due to fibromuscular dysplasia (3). When considering this patient's age at onset of hypertension, and assuming she has renal artery stenosis, a probability of 0.63 of having atherosclerotic disease has been assumed.

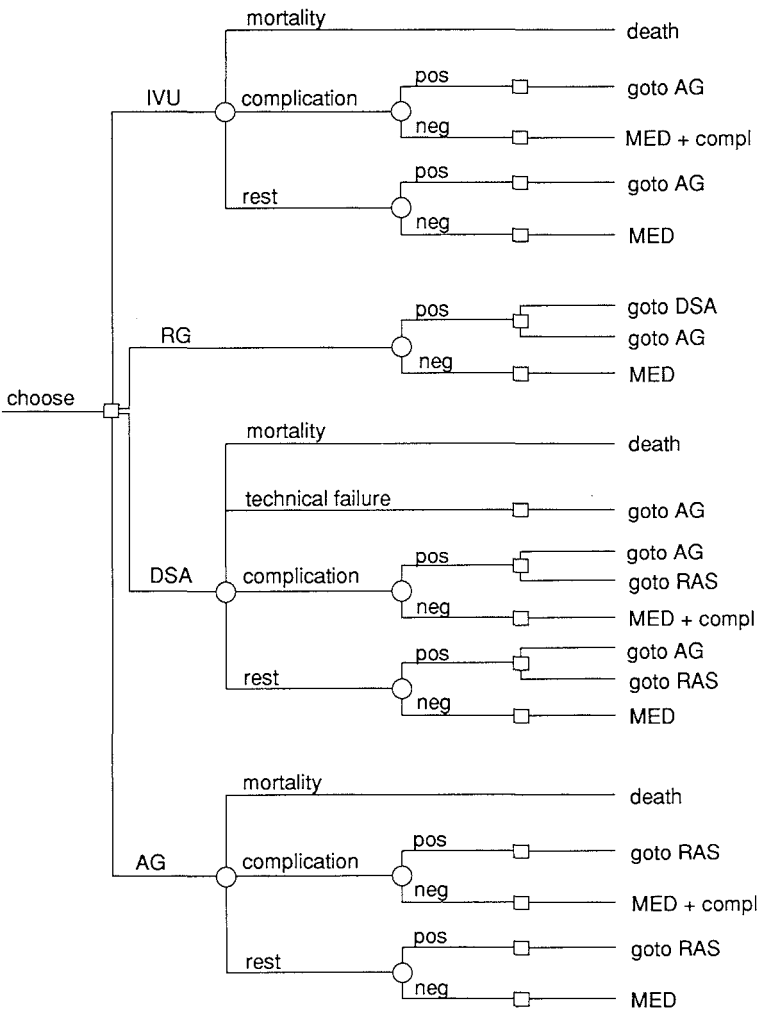
2. Sensitivity and specificity of diagnostic tests:

Arterial angiography is considered the gold standard for the diagnosis of renal artery stenosis. The sensitivity and specificity of other techniques have been determined by taking arterial angiography as the standard. The diagnostic tests considered are assumed to be conditionally independent.

Intravenous urography has a sensitivity of 78% and a specificity of 89% (4,5). The average sensitivity of renography is 86% and the average specificity is 82% (4,6). The average sensitivity and specificity of intravenous digital subtraction angiography is 89% and 93% respectively (7,8,9). Technical failures, due to intestinal movement, a restless patient, or inaccessible veins occur on average in 10% of examinations (7,9,10).

Measuring renal vein renin levels has been advocated prior to doing angiography. However, there is a wide variation in results (11). For this reason the measurement of renal vein renin levels has not been included as a separate diagnostic test.

Figure 1. The decision tree. A square denotes a decision node, a circle denotes a chance node. Further workup or treatment is instituted if the performed test is positive for the diagnosis. Abbreviations used are: IVU = intravenous urography, RG = renography, DSA = intravenous digital subtraction angiography, AG = arterial angiography, RAS = renal artery stenosis, MED = continued medical therapy, PTA = percutaneous transluminal angioplasty, goto = goto subtree, compl = complications, pos = positive test result for renal artery stenosis, neg = negative test result for renal artery stenosis.



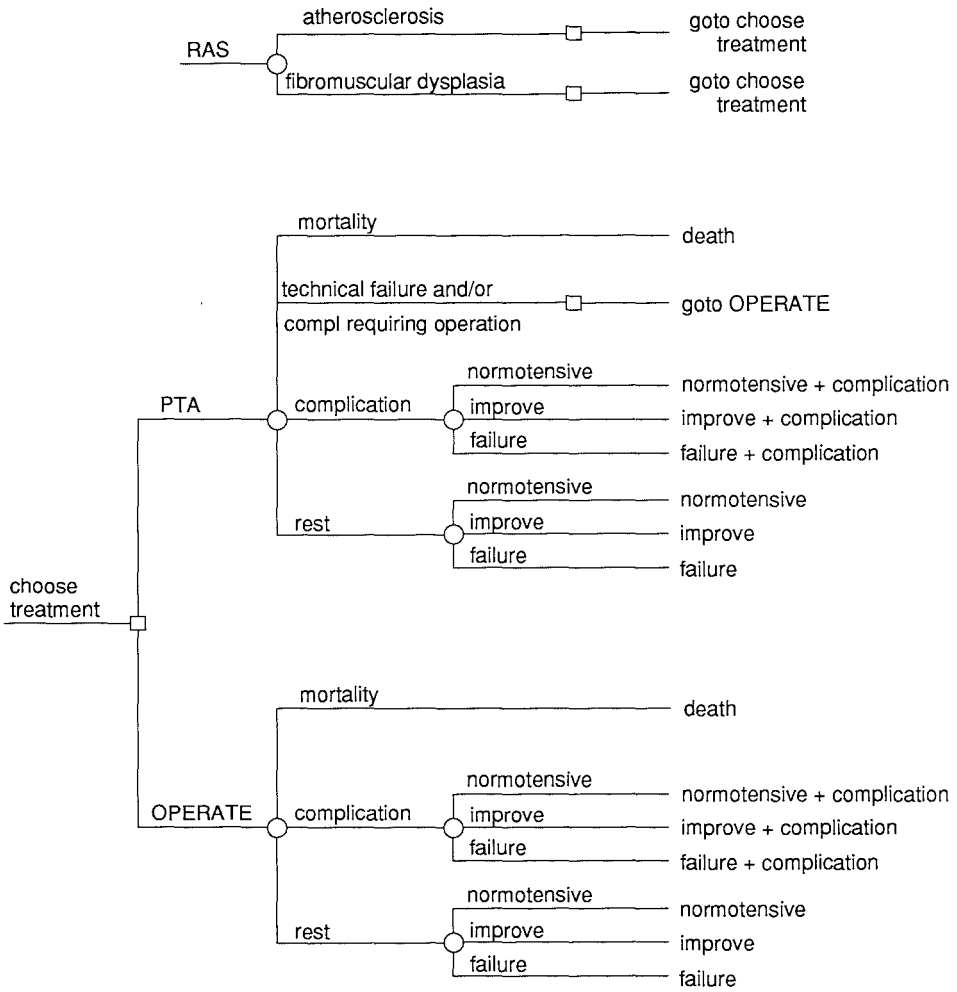

RAS
atherosclerosis
goto choose treatment
fibromuscular dysplasia
goto choose treatment
choose treatment
PTA
mortality
death
technical failure and/or compl requiring operation
goto OPERATE
complication
normotensive
improve
failure
normotensive + complication
improve + complication
failure + complication
rest
normotensive
improve
failure
normotensive
improve
failure
OPERATE
mortality
death
complication
normotensive
improve
failure
normotensive + complication
improve + complication
failure + complication
rest
normotensive
improve
failure
normotensive
improve
failure

Table 1. Assumed probabilities and mortality rates for the case patient. (For abbreviations see legend of figure 1.)

Event	Value	References
prior probability	0.03	1
ratio atherosclerosis: fibromuscular dysplasia	0.63	3
DIAGNOSTIC TESTS		
mortality DSA/IVU	0.00001	12
angiography	0.0003	14
complications DSA/IVU	0.0003	12
angiography	0.017	14
technical failures DSA	0.10	7,9,10,13
sensitivity IVU	0.78	4
renography	0.86	4,6
DSA	0.89	7,8,9
specificity IVU	0.89	4
renography	0.82	4,6
DSA	0.93	7,8,9

TREATMENT		
operative mortality rate		
atherosclerosis	0.045	3,15
fibromuscular dysplasia	0.012	3,15
operative complication rate	0.13	15
operative results:		
normotensive/improve/failure		
atherosclerotic stenosis	0.40/0.33/0.27	3
fibromuscular stenosis	0.57/0.25/0.18	3
PTA mortality rate	0.004	17-21
PTA complication rate	0.04	17-21
PTA technical failure rate	0.09	17-21
PTA results:		
normotensive/improve/failure		
atherosclerotic stenosis	0.21/0.47/0.32	17-21
fibromuscular stenosis	0.45/0.47/0.08	17-21
AGE and SEX SPECIFIC MORTALITY		
woman 55y, annual average mortality rate	0.037/y	23
EXCESS MORTALITY RATES		
normotensive	0	16
diastolic 100 mmHg (MED)	0.0025/y	16
diastolic 110 mmHg (improve)	0.0045/y	16
diastolic 140 mmHg (failure)	0.0105/y	16
antihypertensive medications	0.0025/y	see text
complications treatment	0.06	16
complications workup	0.03	16

3. Complications of diagnostic tests:

The only serious complication of intravenous urography is adverse contrast reactions. In a survey at the Mayo Clinic 1 in 75 000 patients undergoing intravenous urography had a fatal reaction and 1 out of 3 000 had a life-threatening reaction (12).

The major complications of intravenous digital subtraction angiography include adverse contrast reactions and extravasation of contrast. The incidence of contrast reactions of intravenous digital subtraction angiography is assumed to be the same as that of an intravenous urography. Authors agree that morbidity associated with intravenous digital subtraction angiography is rare and of no significance (10,13).

Major complications of conventional angiography include adverse contrast reactions, and vascular, cardiac and neurological complications. The most commonly used transfemoral approach for renal angiography has a mortality rate of 3 in 10 000 and a morbidity rate of 17 in 1000 patients examined (14).

4. Surgery:

Surgical intervention implies aortorenal bypass, endarterectomy with or without patch grafts, end-to-end anastomosis or nephrectomy.

The assumed probability of operative mortality is the weighted average of the mortality rates of various studies: 0.045 for patients with atherosclerotic disease and 0.012 for those with fibromuscular dysplasia (3,15). Patients with a history of angina pectoris and/or myocardial infarction or an impaired renal function have a higher operative mortality rate of up to 23 % (15).

The diastolic blood pressure is used as a yardstick for therapeutic success (16). The criteria used to evaluate the results of surgical intervention are as follows:

"Cure or normotensive" is a diastolic blood pressure of at most 90 mmHg with a decrease (from the pre-treatment diastolic pressure) of at least 10 mmHg, without using antihypertensives. "Improved" is a diastolic pressure of between 90 and 110 mmHg with a drop of at least 15% or normotensive with drug therapy. Others are regarded as "failures" (3).

In this paper all results are expressed in percentages of post-operative survivors cured, improved or not benefitted. Cure rates for fibromuscular dysplasia are higher than those for atherosclerotic disease.

5. Percutaneous transluminal angioplasty (PTA):

A technically successful balloon dilatation is defined as a post-dilatation stenosis of less than 50% . The mortality rate of the procedure is about 0.4% and the complication rate about 4 %

(17-21). Reported technical failure rates vary from 4 % to 33 % (17-21). In the case presented a value of 9 % is assumed for the technical failure rate (weighted average). Higher failure rates have been reported in atherosclerotic lesions, especially in bilateral and ostial or occlusive stenoses (21).

Clinical results are defined in the same manner as post-operative results. As in the case of surgery cure rates for fibromuscular dysplasia are higher than those for atherosclerotic disease.

IV. ASSIGNMENT OF UTILITIES

1. The DEALE-method:

Life expectancy was estimated by the declining exponential approximation of life expectancy (the DEALE-method) (22). This method assumes that life expectancy can be approximated on the basis of a total average constant annual mortality rate. The total mortality rate is the sum of the annual (baseline) mortality rate due to causes other than hypertension and excess mortality rates due to hypertension and procedural complications.

The utility for the outcome "normotensive" for a woman of 55 years is assumed to be her life expectancy as found in the appropriate life table based on Dutch mortality statistics (23). The annual baseline mortality rate, for a woman of 55 years without hypertension, is derived from this life expectancy by the DEALE method. For the outcomes in which the patient remains hypertensive an excess mortality rate due to hypertension is added to the baseline mortality rate. If a procedural complication occurs an excess mortality rate due to the complication is added. From the total average annual mortality rate thus obtained, the life expectancy is calculated. (See appendix I for a sample calculation. Similar calculations have been performed for every outcome.)

The patient is assumed to be risk neutral and no adjustments for quality of life have been made.

2. The excess mortality rates:

A summary of the most important excess mortality rates used in this analysis is given in table I.

The main risk of hypertension is cardiovascular disease. The excess mortality rates for hypertension have been derived from the Framingham Study (16). In case of a negative test result, and thus continued medication, the already achieved blood pressure with initial medications is assumed to remain constant with time and is used as the basis for the calculation of life expectancy.

The use of antihypertensive drugs carries a certain risk in itself. Based on results of trials on antihypertensive drugs the excess mortality rate due to the use of antihypertensive drugs was estimated to be of the order of 0.0025 per patient year (24).

Major complications of therapeutic or diagnostic intervention are assumed to be mainly cardiovascular and neurological disease. The cardiovascular events probably occur about twice as often as the neurological (14,18,19,21). The excess mortality rates of these events have been derived from a review on mortality rates (16). The excess mortality rate of complications caused by a diagnostic test is assumed to be about half of the excess mortality rate of complications caused by therapeutic intervention.

V. RESULTS

1. Calculating the expected utilities of the strategies:

Table 2. Expected utility of various strategies expressed in expected life years for the case patient using the values listed in table I.

Strategy	Expected utility in life years
renography-DSA-PTA	23.66
DSA-angiography-PTA	23.65
IVU-angiography-PTA	23.64
renography-angiography-PTA	23.63
DSA-PTA	23.62
angiography-PTA	23.51
angiography-operate	23.49
no test-PTA	23.05
proven atherosclerotic stenosis:	
PTA	23.40
operate	21.83
proven fibromuscular stenosis:	
PTA	24.82
operate	23.25

The expected utilities are calculated by folding back (2). The principle of folding back is simply taking the sum of the values of all possible outcomes, weighting each outcome value for the probability that the outcome will occur. (For a sample calculation see Appendix II.)

Folding back for the values assumed for the case patient (Table 1) we derive the expected utilities in terms of life years. The expected utilities of various strategies are given in table 2. The expected utility of the strategy "renography - if positive do intravenous digital subtraction angiography - if positive do angioplasty" is the highest, that is 23.66 life years. Strictly speaking the strategy "no test - angioplasty" is a theoretical option equivalent to doing selective renal angiography immediately followed by angioplasty if a stenosis is present.

In the case of a proven stenosis, atherosclerotic or fibromuscular dysplastic, angioplasty has a higher expected utility than surgery.

2. Sensitivity analysis:

To determine the effect of uncertainty in the probabilities assumed we performed a sensitivity analysis. Sensitivity analysis is the calculation of the expected utilities for different values of a particular probability. A decision threshold is the value of the probability below which one strategy will be the best choice and above which another strategy will be the best choice (2).

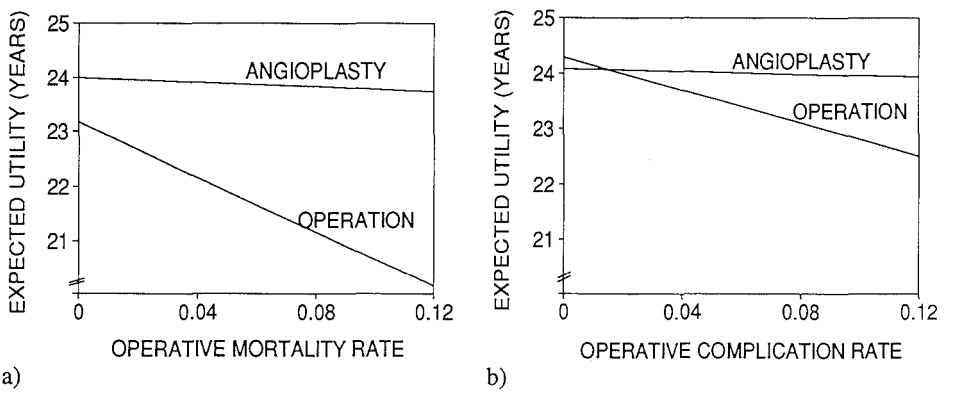


Figure 2. One-way sensitivity analysis for the treatment strategies varying a) the operative mortality rate and b) the operative complication rate. Note that the expected utility of angioplasty depends in part on the utility of surgery because if angioplasty fails technically, surgery will follow.

Figure 2a gives a sensitivity analysis for the operative mortality. Even for very low operative mortality rates angioplasty is better than surgery. The expected utility of angioplasty depends to a small extent on the utility of surgery because if angioplasty fails technically, surgery will follow. Figure 2b shows the results of a sensitivity analysis of the operative complication rate. For very low operative complication rates, that is less than 0.015, surgery will be the treatment of choice, otherwise angioplasty is preferable. These sensitivity analyses were repeated for atherosclerosis and fibromuscular dysplasia separately. The same general trends were found.

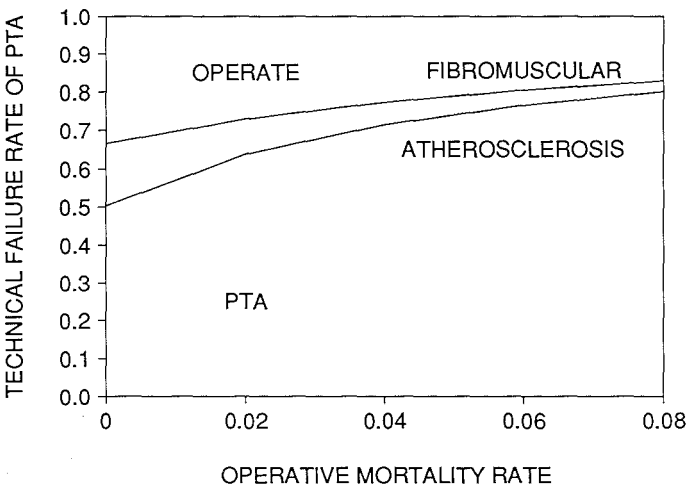


Figure 3. Two-way sensitivity analysis for the treatment options PTA and surgery varying operative mortality rate and technical failure rate of PTA, done for fibromuscular dysplasia and atherosclerotic stenoses separately. The areas marked "OPERATE" and "PTA" denote where surgery and PTA respectively are the treatment of choice.

Figure 3 gives the results of a two-way sensitivity analysis for the technical failure rate of angioplasty and operative mortality. The threshold values of the technical failure rate of angioplasty are given as a function of the operative mortality rate. The analysis has been done for both atherosclerotic stenosis and fibromuscular dysplasia separately. Each line divides the plot into two areas: above the line the combinations of operative mortality and technical failure rate of angioplasty are such that surgery is preferred. Below the line angioplasty is the treatment of choice.

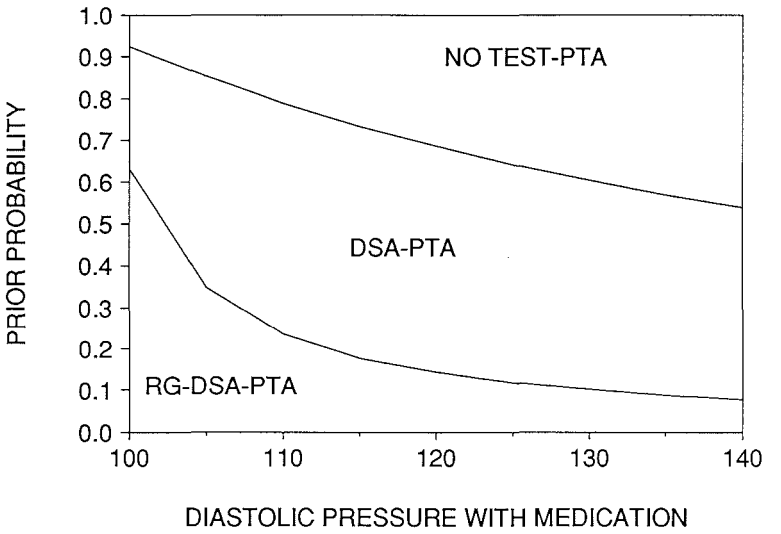


Figure 4. Two-way sensitivity analysis for the overall strategy varying the diastolic blood pressure under initial medical therapy and the prior probability for a renal artery stenosis. (For abbreviations see legend of fig 1.)

A sensitivity analysis for the prior probability of renal artery stenosis has been done. For a prior probability below 0.63 the best strategy is "renography - if positive do DSA - if positive do angioplasty". For a prior probability in the range of 0.63 to 0.92 the best strategy is "DSA - if positive do angioplasty". If one estimates the diagnosis renal artery stenosis to be virtually certain, that is over 0.92, one should proceed to angioplasty without any further diagnostic test.

The calculated thresholds for a change in strategy depend on the diastolic blood pressure with initial medical therapy. A two-way sensitivity analysis has been done determining the threshold prior probability as function of the diastolic blood pressure with initial medication, the results of which are presented in figure 4. The graph is divided into three areas. For the case patient the relevant point is in the bottom left corner in the area where the strategy "renography - if positive do DSA - if positive do angioplasty" is the best choice. It should be borne in mind that this graph has been calculated as an example and applies to a 55 year old Dutch woman. A new graph should be calculated for other ages and/or sex.

VI. COMMENT

The purpose of decision analysis is to rank possible clinical strategies in order of decreasing utility. The important point is the ranking of the options and not the absolute values of the utilities. Rather than choosing arbitrary units life expectancies have been used in this analysis.

The calculated life expectancies are inaccurate for two reasons. The first reason is the fact that the DEALE method yields an approximation of life expectancy which underestimates life expectancy compared with the Gompertz model (22). However, because the age and sex specific mortality is used as starting value for the calculations of ALL outcomes, a slight inaccuracy in the age and sex specific mortality will affect all outcomes alike. The absolute value of the expected utilities is slightly incorrect but the ordering of the strategies should not change. This was verified by a downward adjustment of the age and sex specific mortality which was found not to affect the ordering of the strategies.

The second reason is that the excess mortality rates used are estimates. However, if the rates are correctly ranked the ordering of the preferred strategies will be correct by the same reasoning as above.

No adjustment has been made for the quality of life. If the severity of non-fatal morbidity is assumed to be proportional to the rate of fatal morbidity an adjustment for the quality of life would affect only the absolute value of the outcomes but will not affect the ranking of the options.

We have assumed the patient to be risk neutral. However this is not always the case. A person can be risk averse or risk seeking (2). To take this into account in a decision analysis one would have to obtain a utility curve of how the patient values survival in the immediate coming years as opposed to later years (2).

The utilities in this analysis are based on the results of the Framingham Study on hypertension. Clinical trials have shown that life expectancy of hypertensives can be increased by decreasing the blood pressure. We assume that the blood pressure decrease attained during therapy can be translated to a life expectancy increase using the Framingham risk function, incorporating the risk of the particular form of therapy involved.

In this analysis information concerning chance events has been derived from different sources. This may raise doubts as to the applicability of the model to a specific setting. In the ideal situation the model would be adjusted for local and situative values of the parameters.

Clinicians' estimates take a part in the decision. Asking the radiologist, for example, to estimate the probability of a successful intravenous digital study and/or a successful angioplasty can be very useful. The estimate can be incorporated into the analysis. The calculations done show that even if the technical failure rate of angioplasty is high, angioplasty will be preferred to surgery.

As was to be expected this analysis shows that the prior probability of renal artery stenosis is of the utmost importance in deciding which workup to perform. With the use of a database, registering patient characteristics of patients seen, it should be possible to determine the prior probability of a stenosis given the patient's age, sex and prior history. Preferably a clinical prediction rule should be derived expressing the prior probability as a function of relevant patient characteristics for the patient population concerned.

VII. REFERENCES

1. Thornbury JR, Stanley JC, Fryback DG: Optimizing work-up of adult hypertensive patients for renal artery stenosis. RCNA 1984; 22: 333-339.

2. Weinstein MC, Fineberg HV. Clinical Decision Analysis. Philadelphia, Saunders, 1980.

3. Stanley JC. Renovascular Hypertension: the surgical point of view. In: Clinical aspects of renovascular hypertension (Schilfgaarde R van,ed). Boston: Martinus Nijhoff 1983; 259-268.

4. McNeil BJ, Varady PD, Burows BA, Adelstein SJ: Cost- effectiveness calculations in the diagnosis and treatment of hypertensive renovascular diasease. N Engl J Med 1975; 293: 216-221.

5. McNeil BJ, Adelstein SJ: The value of case finding in hypertensive renovascular disease. N Engl J Med 1975; 293: 221-226.

6. Gruenewald SM, Collins LT: Renovascular Hypertension: Quantitative Renography as a Screening Test. Radiology 1983; 149: 287-291.

7. Buonocore E, Meaney TF, Borkowski GP, Pavlicek W, Gallagher J: Digital Subtraction Angiography of the Abdominal Aorta and Renal Arteries: Comparison with Conventional Aortagraphy. Radiology 1981; 139: 281-286.

8. Smith CW, Winfield AC, Price RR, Harding DR, Tucker SW, Witt WS, Hollifield JW: Evaluation of Digital Venous Angiography for the Diagnosis of Renovascular Hypertension. Radiology 1982; 144: 51-54.

9. Clark RA, Alexander ES: Digital subtraction angiography of the renal arteries: prospective comparison with conventional arteriography. Invest Radiol 1983; 18: 6-10.

10. Hillman BJ: Digital Radiology of the Kidney. RCNA 1985; 23.2: 211-226.

11. Geller SC. Renovascular Hypertension. In: Brigham and Women's Hospital Handbook of Diagnostic Imaging (McNeil BJ, Abrams HL, eds.) 1st ed. Boston: Little, Brown 1986; 77-82.

12. Hartman GW, Hattery RR, Witten DM, Williamson Jr. B.: Mortality during excretory urography: Mayo Clinic experience. AJR 1982; 139: 919-922.

13. Hillman BJ, Ovitt TW, Capp MP, Fisher HD, Frost MM, Nudelman S: Renal Digital Subtraction Angiography: 100 cases. Radiology 1982; 145: 643-646.

14. Hessel SJ, Adams DF, Abrams HL: Complications of Angiography. Radiology 1981; 138: 273-281.

15. Franklin SS, Young JD, Maxwell MH, Foster JH, Palmer JM, Cerny J, Varady PD: Operative Morbidity and Mortality in Renovascular Disease. JAMA 1975; 231: 1148-1153.

16. Singer RB, Levinson L: Medical Risks. 1st ed. Lexington MA: DC Health and Co 1976; 2/10-3/80.

17. Martin LG, Price RB, Casarella WJ, Sones PJ, Wells JO, Zellmer RA, Chuang VP, Silbiger ML, Berkman WA: Percutaneous Angioplasty In Clinical Management of Renovascular Hypertension: Initial and Long-Term Results. Radiology 1985; 155: 629-633.

18. Martin LG, Casarella WJ, Alspaugh JP, Chuang VP: Renal Artery Angioplasty: Increased Technical Success and Decreased Complications in the Second 100 Patients. Radiology 1986; 159: 631-634.

19. Tegtmeyer CJ, Kellum CD, Ayers C: Percutaneous Transluminal Angioplasty of the Renal Artery. Results and Long-Term Follow-up. Radiology 1984; 153: 77-84.

20. Boomsma JHB: Percutaneous Transluminal Dilatation of Stenotic Renal Arteries in Hypertension. The Dotter technique as applied to the renal artery. Thesis. Groningen: van Denderen BV 1982.

21. Sos TA, Pickering TG, Sniderman K, Saddekni S, Case DB, Silane MF, Vaughan Jr ED, Laragh JH: Percutaneous Transluminal Renal Angioplasty in Renovascular Hypertension due to atheroma or fibromuscular dysplasia. N Engl J Med 1983; 309: 274-279.

22. Beck JR, Kassirer JP, Pauker SG, Gottlieb JE: A convenient approximation of life expectancy (the DEALE) I. Validation of the method. II. Use in medical decision making. Am J Med 1982; 73: 883-897.

23. Central Bureau for Statistics/ Ministry of Health: Vademecum: Health Statistics of the Netherlands 1983. 's- Gravenhage, Staats Uitgeverij 1983.

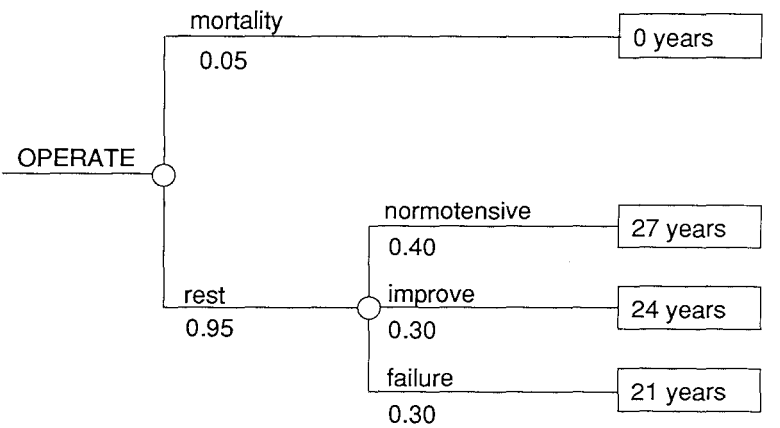
24. MacMahon SW, Cutler JA, Furberg CD, Payne GH. The effects of drug treatment for hypertension on morbidity and mortality from cardiovascular disease: a review of randomized controlled trials. Prog Cardiovasc Dis 1986; 29-3 suppl 1: 99-118.

APPENDIX 1.

As described in the DEALE-method we have used the age and sex specific life expectancy (for a healthy person) and the excess mortality rates (for the diseases the patient has) to calculate life expectancies for the various outcomes (22). As an example the calculation of the life expectancy of the outcome "MEDICATION" (without morbidity from workup or intervention) is given here.

1) average life expectancy 55 year old woman (life tables the Netherlands (23))	= 27.0 y
2) average annual baseline mortality rate (inverse of (1))	= 0.037/y
3) excess mortality rate for diastolic blood pressure of 100mmHg, woman of 55y (table I)	= 0.0025/y
4) excess mortality rate for use of anti-hypertensive medications (table I)	= 0.0025/y
5) total annual mortality rate for outcome (sum of (2),(3) and (4))	= 0.042/y
6) life expectancy for outcome (inverse of (5))	= 23.8 y

APPENDIX 2.



As an example folding back for the choice "operate" in the case of proven atherosclerotic stenosis is presented here. The subtree has been simplified by ignoring the possibility of complications and by using rounded figures.

The outcome value of every branch is weighted for the probability that the outcome will occur. For the branch "normotensive" the weighted outcome is 0.4 x 27 = 11 life years, for the branch "improve" 0.3 x 24 = 7 and the branch "failure" 0.3 x 21 = 6 life years. The expected utility of the branch "rest" is the sum of the weighted values of its branches which is 11 + 7 + 6 = 24 life years. In the case of operative mortality the outcome value is zero life years. The sum of the weighted values of the branches of the choice "OPERATE" is (0.05 x 0) + (0.95 x 24) giving an expected utility of 23 life years.

Chapter IV

PERCUTANEOUS NEPHROSTOMY FOR ACUTE URINARY TRACT OBSTRUCTION CAUSED BY UROLITHIASIS[1]

I. INTRODUCTION
II. THE PROBLEM STATED
III. SUMMARY OF AVAILABLE DATA
1. Natural history of ureter stones
2. Ultrasound for urinary tract obstruction and pyonephrosis
3. Percutaneous nephrostomy (PCN)
4. Retrograde ureteral stenting (RUS)
5. Surgical intervention in pyonephrosis
6. Excess mortality rates
7. Definitive treatment for a ureteral stone
IV. THE MODEL
V. TECHNICAL DETAILS
VI. ASSUMPTIONS
VII. RESULTS
1. The best strategies
2. Probability of pyonephrosis and obstruction at presentation
3. Ultrasound examination and dilatation of the collecting system
4. Mortality due to sepsis and the probability of developing sepsis
5. Solitary kidney
6. Increased bleeding tendency and increased risk of PCN/RUS
7. Size and position of the stone
8. Definitive treatment and timing of intervention
VIII. DISCUSSION
IX. REFERENCES

ABSTRACT

Obstruction of the urinary tract, often caused by urolithiasis, may lead to pyonephrosis, sepsis, renal insufficiency and death. Consequently, to relieve the obstruction a temporary drainage procedure, namely percutaneous nephrostomy or retrograde ureteral stenting, may be performed.

---

[1]Co-author: John Wong, Division of Clinical Decision Making, Department of Medicine, New England Medical Center, Tufts University Medical School, Boston.

These procedures, however, carry a risk of haemorrhage or infection. Moreover, with non-invasive medical management, spontaneous passage of the stone may occur making any intervention superfluous. The likelihood of spontaneous passage depends mostly on the size and position of the stone. The likelihood of complications from the procedures depends on the presence or absence of infection, dilatation of the urinary tract and risk factors for increased bleeding. Many of the probabilities involved in the analysis depend on the duration of symptoms and/or obstruction, which may be conveniently modelled with a Markov process. We examined the trade-off between medical management, percutaneous nephrostomy or retrograde ureteral stenting with a Markov process decision tree.

I. INTRODUCTION

Occurring in 3.5 to 3.8% of all autopsies, urinary tract obstruction is common and is a remediable cause of kidney failure (20). Over half of the cases of urinary tract obstruction are caused by urolithiasis (70). Other causes include fibrosis from previous surgery or radiation, often for malignant disease, especially in the small pelvis. Bilateral obstruction is usually associated with loss of renal function and significant morbidity. For example, 60 to 90 % of patients with cervical cancer die of uraemia if left untreated (8,44,62). In this paper we will focus on urolithiasis as cause of urinary tract obstruction. When associated with infection of the tract, which is not infrequent, it is defined as pyonephrosis. Timely diagnosis and appropriate treatment are necessary to prevent gram-negative sepsis, loss of renal function or death.

Left untreated, patients with urolithiasis may develop sepsis and/or renal failure which may be fatal. Consequently, these patients have, in the past, been treated surgically with nephrectomy. Because of the associated high mortality and morbidity rate, more recently less invasive procedures, such as percutaneous nephrostomy (PCN) and retrograde ureteral stenting (RUS), are usually performed. The risks of percutaneous nephrostomy include haemorrhage and infections, either exacerbation of existing infection or introduction of infection. A lower chance of successful placement of percutaneous nephrostomy is observed in non-dilated collecting systems. The risks of retrograde ureteral stenting include injury to the ureter or renal pelvis and exacerbation of existing infection.

The decision whether to drain the tract immediately, by either PCN or RUS, or delay intervention and observe the patient is not clear-cut. Although non-invasive medical management may result in sepsis and/or renal failure, the stone causing obstruction may pass spontaneously, obviating the need for intervention. An expectant approach may also result in increased dilatation of the ureter, facilitating performance and increasing the likelihood of a successful nephrostomy.

Consider the following four cases:

1.
A 40 year old man with a small stone in the lower ureter has fever, bacteriuria, haematuria and flank pain, suggesting pyonephrosis. On ultrasound, dilatation of the urinary tract is seen, however, no echogenic material is identified within the tract.

2.
A 65 year old man with a history of urolithiasis presents with urosepsis and urinary tract obstruction, probably caused by a ureteral stone. However, stones are not seen on the plain film of the abdomen. An ultrasound reveals mild dilatation of the collecting system and the presence of echogenic material within the tract cannot be judged reliably. The patient has a low platelet count and a prolonged bleeding time.

3.
A 60 year old man with a solitary kidney and a staghorn calculus, one day post extracorporeal shockwave lithotripsy (ESWL), develops anuria, flank pain and urosepsis, suggesting acute obstruction, but has no evidence of dilatation of the urinary tract.

4.
A 40 year old man presents with since a few days complaints of colicky flank pain, with neither fever nor pyuria. A plain film of the abdomen shows a small stone in the lower ureter. An ultrasound reveals dilatation of the tract, but echogenic material is not identified in the collecting system.

Should each of these patients have a percutaneous nephrostomy or retrograde stenting performed or should intervention be delayed? The answer is not clear-cut and involves many uncertainties. The trade-offs include the trade-off between the risk of the procedure, which increases with an increased bleeding tendency (case 2) or a solitary kidney (case 3), and the risk of medical management, which increases with a large stone (case 3), obstruction (all the cases), pyonephrosis (case 1) and sepsis (cases 2 & 3). Another trade-off is that between the benefit of medical management, in that the stone may pass spontaneously (cases 1,2,4), and the benefit of drainage in preventing death from sepsis (cases 1,2,3) and preventing renal impairment (all the cases). Although algorithms have been developed and general indications for percutaneous nephrostomy have been published (22), a controlled trial comparing drainage of the tract by percutaneous nephrostomy or retrograde stenting and medical treatment has not been published. Moreover, these general guidelines do not consider individual risks and benefits.

Decision analysis is an explicit approach to making decisions in the presence of uncertainty. With decision analysis we can compare treatments and analyze clinical practices. In this paper we present a decision analytic approach to the treatment of urinary tract obstruction caused by urolithiasis, focusing on whether a drainage procedure should be done and, if so, which one (PCN or RUS) and when.

II. THE PROBLEM STATED

Should a percutaneous nephrostomy (PCN) or retrograde ureteral stenting (RUS) be performed in patients with urinary tract obstruction caused by urolithiasis? Medical management of urinary tract obstruction may lead to sepsis, renal insufficiency or death, but the obstruction may also spontaneously relieve itself by passage of the stone. A percutaneous nephrostomy or retrograde stenting procedure relieves the obstruction acutely but may be associated with haemorrhage or infection. We examine in detail the influence of pyonephrosis, sepsis, an increased bleeding tendency, a non-dilated tract and/or a solitary kidney. Furthermore, we examine the effects of the presence or absence of obstruction. We focus on whether a procedure is necessary and, if so, when it should be performed. The time horizon of the analysis is restricted to the acute problem of urinary tract obstruction.

Table 1. Summary (and abbreviations) of the strategies. The analyzed strategies are combinations of an initial treatment and an alternative treatment if the initial strategy fails or if obstruction recurs. Two diagnostic strategies are examined for cases in which ultrasound has not yet been performed.

initial treatment	alternative treatment if initial strategy fails	abbreviation of strategy
percutaneous nephrostomy (PCN)	PCN	PCN-PCN
	RUS	PCN-RUS
	OPER	PCN-OPER
	MED	PCN-MED
retrograde ureteral stenting (RUS)	PCN	RUS-PCN
	RUS	RUS-RUS
	OPER	RUS-OPER
	MED	RUS-MED
surgical management		OPER
medical management (MED)	MED	MED
	if sepsis occurs, PCN	MED-sep-PCN
	if sepsis occurs, RUS	MED-sep-RUS
	if pyonephrosis/sepsis occur, PCN	MED-pyo-PCN
	if pyonephrosis/sepsis occur, RUS	MED-pyo-RUS
	if pyonephrosis/sepsis occur, or if stone has not passed after 15 days, PCN	MED-time-PCN
	if pyonephrosis/sepsis occur, or if stone has not passed after 15 days, RUS	MED-time-RUS
if dilated on US, PCN if non-dilated MED	MED	USdil-PCN
	US every 5 days, if dilatation occurs, PCN	MED-USdil-PCN

We examine the following strategies: PCN (percutaneous nephrostomy), RUS (retrograde ureteral stenting), surgical intervention, and medical management, with or without a procedure performed conditional on the development of complications. The strategies analyzed are combinations of an initial treatment and an alternative if the initial strategy fails or if obstruction recurs. Table 1 summarizes the 18 strategies. The initial treatment may be PCN, RUS, surgery or medical management. If an initial drainage procedure fails, or obstruction recurs, the initial procedure may be repeated or another treatment may be instituted (table 1). The decision is examined partly with a Markov process, each Markov cycle representing 5 days, during which two interventional procedures may be chosen. If an interventional strategy is chosen and both procedures are unsuccessful, the patient will undergo surgery.

III. SUMMARY OF AVAILABLE DATA

1. Natural history of ureter stones

a) Spontaneous passage of ureter stones:

In 1956, Sandegard details the natural history of ureteral stones demonstrated on radiographic films (55). 324 patients were followed until spontaneous passage of the stone occurred or active intervention was indicated because of anuria, sepsis or gross anatomic changes requiring surgery. Medical treatment consisted of analgesics and spasmolytics. The data includes the number of cases of spontaneous passage at 1, 2, 4, and 12 weeks and at 18 months after onset of symptoms, and the likelihood of passage is stratified by size, position, shape and primary versus recurrent stones. Thus, the data provides the overall cumulative probability of passage and the rate of passage over time. Overall, spontaneous passage occurred in 73% of cases. We modelled the probability of spontaneous passage depending on size and position of the stone because these are the two main features that determine the likelihood of spontaneous passage. Although other features increase the likelihood of spontaneous passage, they are usually also associated with either smaller size or lower position.

The size of the stone is determined on radiographic films taken under standardized conditions with a focus-film distance of one meter. About 90% of ureteral stones contain enough calcium to be visible on a plain film of the abdomen (65). Stones are classified by their width: less than 4 mm is small, more than or equal to 4 mm but less than 6 mm is medium, and more than or equal to 6 mm is large. The position in the ureter is classified as "upper" if the stone is above the level where the ureter crosses the iliac vessels and otherwise it is classified as "lower". Thus, depending on size and position, there are six types of stones. For each of these six groups we calculated the observed cumulative probability of spontaneous passage of the stone in relation

to the time after onset of symptoms. Subsequently we modelled the cumulative probability (or cumulative incidence) of spontaneous passage as a function of time using a Weibull fit (36), of the form:

$$pass(t) = c\left(1 - e^{-(\lambda t)^{\gamma}}\right) \qquad (Eq.1)$$

where c is a constant equal to the cumulative probability of passage for $t \rightarrow \infty$, γ determines the shape of the curve and λ is a scaling constant. The corresponding hazard function h(t) is:

$$h(t) = \frac{c\lambda\gamma(\lambda t)^{\gamma-1}}{(1-c)e^{(\lambda t)^{\gamma}} + c} \qquad (Eq.2)$$

Theoretically, other shapes of the curve are possible. The exponential model:

$$pass(t) = c\,(1 - e^{-rt}) \qquad (Eq.3)$$

would be an alternative form, however this model assumes a constant hazard function or rate r. The Weibull model is a generalization of the exponential model in which the hazard function may vary with time. Note that if $\gamma = 1$ the Weibull model is equivalent to the exponential model. Although the exponential model is computationally easier, our Weibull model has a lower chi-squared statistic[1] and thus provides a better fit (36).

To facilitate determining constants for the Weibull model, we transformed the above equation Eq. 1 as follows[2]:

$$\ln\left(\ln\left(\frac{c}{c - pass(t)}\right)\right) = \gamma\ln(\lambda) + \gamma\ln(t)$$

Substituting:

$y = \beta_1 + \beta_2 x$

$$y = \ln\left(\ln\left(\frac{c}{c - pass(t)}\right)\right)$$

---

[1]The test statistic for the chi-squared goodness-of-fit test equals:

$$\chi^2 = \sum_k \frac{(O-E)^2}{E}$$

with k-r-1 degrees of freedom, where O are the observed values, E the expected values on the basis of the regression equation, k the number of data points and r the number of estimated parameters.

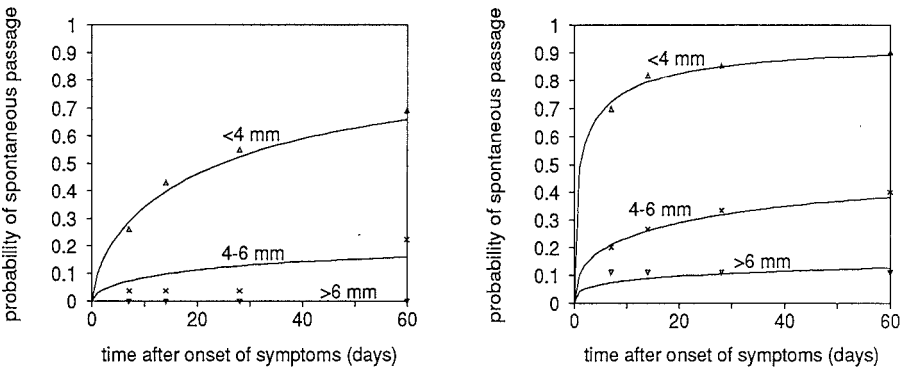
[2]Note: ln(x) is the natural logarithm of x

$x = \ln(t)$

$\beta_1 = \gamma \ln(\lambda)$

$\beta_2 = \gamma$

yields a linear equation with independent variable x, permitting a least squares regression to estimate γ and λ.



a) b)

Figure 1. Cumulative probability of spontaneous passage of a stone as function of time, depending on position and size of the stone. **a)** Stones in the upper ureter and **b)** stones in the lower ureter.

For small (<4 mm), medium (4-6 mm) and large (>6 mm) sized stones, figure 1 presents the observed data points and derived cumulative probability functions in a) the upper and b) the lower ureter. Large stones in the upper ureter never pass spontaneously. Note that the estimated constants, γ and λ, for medium sized stones in the upper ureter and large stones in the lower ureter are based on small samples (27 and 9 cases respectively). To model the probabilities for these two groups, we assumed that the shape of the curve (and thus γ and λ) is the same as that of upper ureter or large stones in general. We then scaled the curve down proportional to the overall probability of passage in this subgroup. (The estimated parameters provide an adequate fit, at the 0.05 level of significance, for all the curves except one: the level of significance for the fitted curve for medium sized stones in the upper ureter is 0.20.)

More recent studies of the natural history of urolithiasis generally agree with the Swedish study. In a study of calcium urolithiasis (9), passage occurred in 64-80% of 460 cases compared to the 73% in the Swedish study. A review (11) summarizes the probability of spontaneous passage within one year of onset of symptoms, for stones located in the lower ureter stratified by size

(Table 2). Compared to the Swedish study, the probability of passage reported is slightly lower for small stones and slightly higher for medium sized and large stones (Table 2). For the analysis we used the probability of passage modelled with the data from the Swedish study.

b) Dilatation:

Not all urinary tracts dilate immediately as a result of obstruction. Among patients presenting with urinary tract obstruction, tracts were not dilated in about 4% of 431 cases at the time of percutaneous nephrostomy (4,38,42,71). However, if obstruction is not relieved, the likelihood of dilatation increases with increasingly prolonged obstruction.

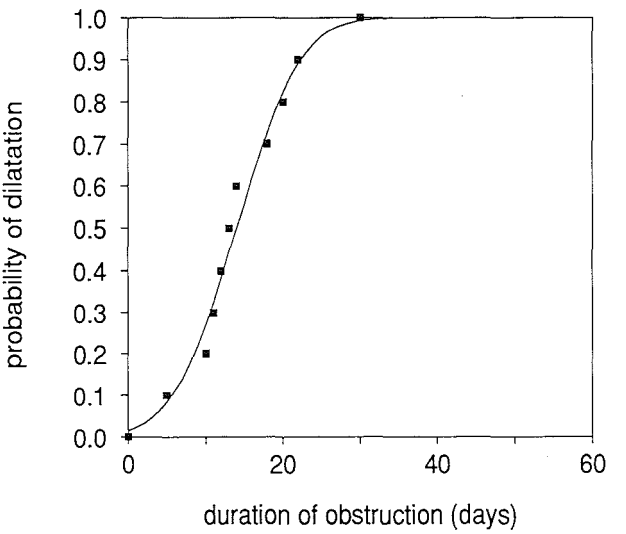


Figure 2. Cumulative probability of dilatation in relation to duration of obstruction, observed values (squares) with fitted curve.

Because data on dilatation of the urinary tract in relation to the duration of obstruction in humans is sparse, we refer to animal experiments. In 10 dogs, incomplete unilateral ureteral ligation led to progressive hydroureteronephrosis (31). Measurements of intrarenal pressure and pelviureteric volume at various time intervals (up to 60 days) after ligation demonstrated that pressure increases slowly with increasing volume. At a critical capacity, however, the pressure rises much more steeply with increasing volume. This critical capacity of the renal pelvis represents a

transition point from accommodation to overdistension. Because the precise volume at which the renal pelvis is defined as dilated is to a certain extent arbitrary, and because of the above relationship between pressure and volume in the animal experiments, we define dilatation (in dogs) as any volume greater than or equal to the critical capacity. The average duration of partial obstruction after incomplete ligation of the dogs' ureters, at which the critical capacity was reached, is 14 days (31). Based on the cumulative probability of dilatation as function of time, we fitted a sigmoid shaped curve to the observed points, of the form:

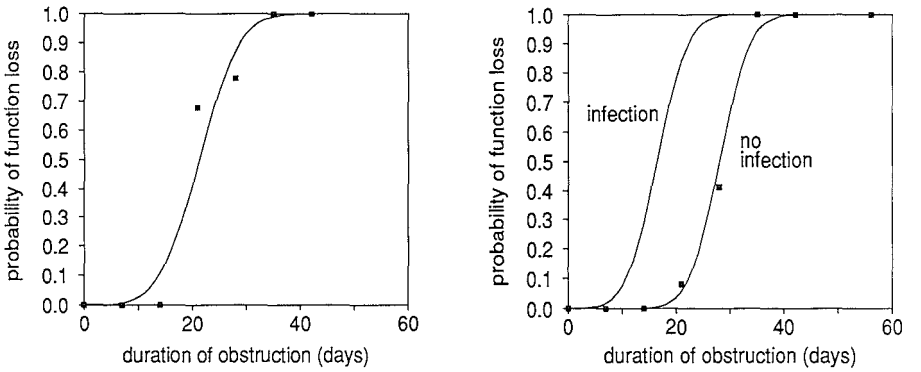
$$dilate(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

where dilate(t) is the cumulative probability of dilatation after t days obstruction, Φ is the standard cumulative normal distribution, μ is the mean time interval after which dilatation occurs (14 days) and σ is the standard deviation of the mean (Figure 2).

c) Loss of renal function:

Obstruction of the urinary tract causes increased pressure in the collecting system, a decreased glomerular filtration rate and hemodynamic changes, which together affect kidney function. If obstruction exists only for a short period of time, functional loss is completely reversible after release of the obstruction (30). Prolonged obstruction results in permanent loss of renal function, in spite of resolution of the obstruction. The likelihood of permanent functional loss increases with the duration of obstruction: the precise relationship, however, is not well known. In addition, infection increases the likelihood of loss of function. However, the presence of a solitary kidney decreases the likelihood of functional loss, presumably due to a compensatory mechanism. Note that defining functional loss is somewhat arbitrary: we assume that a kidney which functions less than 20% of normal is a remnant kidney (39,46).

Studies of renal function in humans after release of acute urinary tract obstruction are sparse. Most studies have focused on chronic obstruction (17,19,29) rather than acute obstruction. Animal experiments in which renal function is measured after varying periods of acute obstruction (27,28,50,51,66) suggest that a sigmoid-shaped curve (that is, a cumulative normal distribution as used in fitting the probability of dilatation above) reflects the biological process (Figure 3a) (the level of significance, using a chi-squared goodness-of-fit test is 0.05). A single study (25) reports prognosis of renal function after release of acute obstruction in 74 patients without infection and 6 patients with infection. Fitting the probabilities given in the study of acute obstruction in man to a sigmoid-shaped curve provides the probability of renal function loss after varying periods of acute obstruction for patients without infection (Figure 3b). The average duration of obstruction after which permanent renal function loss occurred was 28 days for uninfected kidneys and 17.5 days for infected kidneys. Thus, the cumulative probability



a) b)

Figure 3. a) Probability of loss of renal function in animals after release of acute obstruction, depending on the duration of obstruction. Observed values (squares) and fitted curve. **b)** Probability of loss of renal function in humans after release of acute obstruction, depending on the duration of obstruction and presence of infection. Observed values (squares) for patients without infection and fitted curves for patients with and without infection.

curve for patients with infected kidneys is shifted to the left relative to the curve for patients with uninfected kidneys. Data from experiments with rats (8) demonstrate that a solitary kidney can be obstructed approximately 1.25 times as long as a kidney of a pair before it becomes remnant. The curve for solitary kidneys is therefore shifted to the right. Because data for the latter two groups are sparse, we assume that the shape of the curve is the same as for patients with two kidneys without infection, and that infection and/or the presence of a solitary kidney only shifts the curve right or left.

d) Obstruction, pyonephrosis and sepsis:

Among 98 patients presenting with acute flank pain, renal colic and haematuria in the emergency ward, 73 had calculus disease of which 62 had evidence of obstruction on excretory urography. Even if absent initially, obstruction may develop. We modelled the cumulative probability of obstruction given urolithiasis as a time-dependent probability of the form:

$$obstruction(t) = 1 - e^{-rt}$$

where t is the time since onset of symptoms related to the presence of a ureteral stone and r is the rate of developing obstruction per unit of time. Note that by modelling the probability of obstruction with this equation, we assume that, if the stone does not pass spontaneously and is not removed, obstruction will occur eventually. The Swedish study suggests that the probability

of obstruction depends on the size of the stone (55). Over 18 months, obstruction developed in 31% of all patients in whom the stone did not pass. Among patients with small, medium and large stones, 18%, 26% and 45% respectively developed obstruction. We calculated the rate of developing obstruction per day from these percentages. The rate varies from 0.0004 per day for small stones to 0.0011 per day for large stones (table 2).

Among patients presenting with obstruction, about 15% have pyonephrosis (71). If pyonephrosis is suspected clinically because of chills, fever, flank pain and pyuria, the chance of having a positive culture on aspiration is about 86% (71). Not all patients with urinary tract obstruction develop pyonephrosis. The Swedish study suggests that the probability of developing pyonephrosis depends on the size of the stone (55). We modelled the cumulative probability of pyonephrosis for various stone sizes, depending on time, with an exponential failure function with a constant rate, and assuming that obstruction which is not relieved will eventually lead to pyonephrosis:

$$pyonephrosis(t) = 1 - e^{-rt}$$

Among patients with obstruction from a stone, 43% developed pyonephrosis during a follow-up period of 18 months, giving an overall rate of 0.001 per day (55) (table 2). 23% of patients with small or medium sized stones and 64% of patients with large stones developed pyonephrosis, giving rates of 0.0005 and 0.0019 per day respectively. (Note that the rates of developing obstruction and pyonephrosis are very low, which implies that the cumulative probability reaches unity only after a long period of time.)

Among 78 patients with benign urinary tract obstruction and pyonephrosis, 51% presented with sepsis (34). Of 136 patients who acquired urinary tract infection during urological manipulation, 25 died of sepsis (49), from which we infer, assuming that these 25 represent the 38.5% of patients with sepsis that die, that 65 (48%) of the 136 infected patients had sepsis. Thus, on average, the probability of sepsis given infection of the urinary tract, is about 49% (34,49), however, according to expert opinion, the probability could be as high as 70% within hours to days after the onset of pyonephrosis (46). We modelled the cumulative probability of sepsis given pyonephrosis as a time-dependent probability of the form:

$$sepsis(t) = c\,(1 - e^{-rt})$$

where c is the cumulative probability of sepsis and r the rate at which sepsis occurs. A study reports that of the patients who develop bacteriaemia after urological manipulation, 96% had a positive blood culture within two days following the procedure (53), from which we estimate a rate of 1.6 per day.

Even with prompt institution of antibiotics, the probability of death due to gram-negative sepsis ranges from 15 to 51% (on average 38.5%) (13,32,34,56,68), depending largely on the severity of the underlying disease. Fatalities from sepsis usually occur within one to two days after onset of bacteriaemia (33). Performing a drainage procedure decreases the probability of death from

sepsis. The efficacy of drainage refers to the proportion of patients in who death from sepsis is prevented by the procedure. If the efficacy of a procedure is one, the procedure reduces the risk of death to zero. If the efficacy is zero, the procedure does not reduce the risk of death.

2. Ultrasound for urinary tract obstruction and pyonephrosis

Dilatation of the collecting system on ultrasound does not always indicate obstruction of the urinary tract (20). An increased fluid load, vesicoureteric reflux and megacalices are some of the reasons for a false positive ultrasound. In addition, obstruction of the urinary tract may exist without dilatation of the collecting system. However, in clinical practice the decision to intervene is often based on the presence of dilatation on ultrasound examination. We analyzed this approach by including a diagnostic strategy: perform an ultrasound, if dilatation is present do PCN, otherwise manage medically. The patient may also be examined periodically to determine if dilatation has occurred.

Ultrasound performed to diagnose obstruction of the urinary tract has a true positive fraction of 96% (4,38,42,71) and a false positive fraction of 26% (15). Thus, the likelihood ratio positive for the diagnosis obstruction, given a positive ultrasound, is 3.70, and given a negative ultrasound, is 0.05. Ultrasound distinguishes obstruction without infection from pyonephrosis, by identifying echogenic material in the dilated tract, with a true positive fraction of 83% and a false positive fraction of 6% (4,26,63). The likelihood ratio positive for the diagnosis pyonephrosis, given an ultrasound positive for pyonephrosis, is 13.88, and given an ultrasound negative for pyonephrosis, is 0.18. Pyonephrosis is definitively identified by culture of urine obtained directly from the renal pelvis. This requires aspirating urine via a drainage procedure and waiting three days for culture results unless macroscopic pyonephrosis exists as evident by the presence of pus.

3. Percutaneous nephrostomy (PCN)

a) Successful placement of a PCN tube:

On average, 96% of percutaneous nephrostomy tubes are placed successfully (range 92 to 99%) (24,34,42,52,61). In undilated urinary tracts, or in the presence of a staghorn calculus, access to the tract is more difficult and the probability of success is lower, on average 88% (range 85 to 91%) (38,42,52). In addition, an initially unsuccessful placement due to inability to cannulate the tract, lowers the likelihood of subsequent successful placement. Of 9 unsuccessful placements only 3 catheters were successfully placed several days later (24).

b) Mortality and short-term morbidity:

The overall mortality rate for percutaneous nephrostomy is less than 0.2% (47). Direct procedure related mortality is usually due to massive haemorrhage (48,52). Exacerbation of infection is the second most frequent cause of death. The probability of death among patients undergoing PCN for gram-negative sepsis is 8% (34).

Sepsis occurred in 2.7% of 1210 patients undergoing a percutaneous nephrostomy (24,34,42,48,61,71). If infection of the tract exists at the time of PCN, transient bacteriaemia is unavoidable and approximately 11% of patients develop septicaemia (71). Elevation of intrarenal pressure, for example by injecting excessive contrast material, increases the risk of sepsis. In the absence of infection, the likelihood of sepsis is probably lower than the overall observed frequency of 2.7%, however, the exact frequency cannot be found because studies do not report the frequency of sepsis for non-infected patients separately.

Insignificant haemorrhages and haematuria occur frequently after PCN. Of 1285 PCN's, clinically significant haemorrhage occurred in 0.8% of cases (range 0.4% to 2.0%) (24,34,42,48,61,71). Approximately 20% of the clinically significant haemorrhages are serious enough to necessitate either embolization of the renal artery or surgical nephrectomy (24,34,42,48,61,71). If gram-negative sepsis is present, the likelihood of PCN related haemorrhage and death increases because of associated coagulation disorders (33). Procedure related deaths are often due to haemorrhagic complications and usually associated with increased bleeding tendency (48,52). No published data exists on the probability of death if haemorrhage occurs after PCN. Therefore, in patients with a bleeding diathesis we assume that the odds of PCN related mortality increases proportionate to the increase in the odds of haemorrhage.

Inadequate drainage due to dislodgement of the tube, obstruction by debris or blood clots and catheter breakage occurred in 2% of 884 procedures (range 0.6% to 17%) (24,34,61). If the drain stops functioning the procedure is usually repeated.

For patients successfully treated with PCN, the average hospital stay is six days (34). In the event of short-term morbidity we estimate the length of stay to be twice as long.

c) Efficacy of PCN in sepsis:

As mentioned above, the probability of death among patients undergoing PCN for gram-negative sepsis is 8% (34). Among patients with gram-negative sepsis not undergoing a drainage procedure (in the same study) death occurred in 40% (34). The efficacy of PCN (*effPCN*), that is, the proportion of patients in who death from gram-negative sepsis is prevented by a successfully placed nephrostomy tube, is calculated as follows:

$$0.08 = (1 - effPCN) \times 0.40$$

and is therefore 0.80.

d) Renal function:

In 45 cases treated with PCN, 8.8% of patients tested months to years after the procedure, had a remnant kidney (34,71). However, these studies do not document the duration of renal obstruction before drainage. Because duration of obstruction is the major variable that affects renal function, we assume that if a successful drainage is performed immediately, the probability of losing renal function is negligible. In the presence of obstruction, functional loss may occur if intervention is postponed.

4. Retrograde ureteral stenting (RUS)

Retrograde ureteral stenting (RUS) is used in emergency situations to drain the urinary tract, to relieve acute obstruction, and to stabilize the patient until definitive treatment is started (54). It is a relatively safe procedure that is usually performed under local or regional anaesthesia, with general anaesthesia being required in uncooperative patients. Unfortunately, quantitative data about the complications of stenting in acute obstruction due to ureteral stone disease are sparse and most studies combine results of stenting performed for miscellaneous problems, including malignant disease. Stents placed for malignant obstruction may cause additional complications because the stent remains in place for months, causing erosion and being liable to encrustation and obstruction. Ureteroscopy for stone manipulation is a similar procedure to stenting, only slightly more risky because, in addition to passing the stone with a guidewire, stone crushing and stone retrieval instruments are used. Where appropriate and necessary we use data from studies of ureteroscopy as a best estimate for stenting, modified so as not to overestimate the risks.

a) Successful placement of a RUS:

Reported success rates for retrograde stenting, placed for various indications, range from 80 to 82% (11,40). Successful placement of a stent in acute obstruction due to calculus disease depends on whether the stone can be passed with the guidewire and stent, which in turn will depend on the amount of inflammatory oedema and the size and position of the stone. Of 101 ureteroscopies for ureter stone manipulation only 6% were unsuccessful because the stone could not be passed (11). We therefore estimate the success rate of retrograde stent placement for calculus disease to be 94%. Similar to PCN placement, we assume that an unsuccessful initial stent placement

lowers the likelihood of success of a second attempt. We assume that the odds of a successful placement after an initial unsuccessful RUS, will decrease in the same proportion as for a second PCN placement after an initial failure.

b) Mortality and short-term morbidity:

Procedure related mortality due to retrograde ureteral stenting is caused by infection and iatrogenic injury, and in some cases anaesthesia risk. The mortality rate is assumed to be the same as for ureteroscopy, which is 0.5% (11).

The main risks of retrograde stenting include iatrogenic injury to the renal pelvis or ureter and sepsis (3,40,60). Serious complications occur in 3 to 5% of patients undergoing retrograde stenting (3). About half of these are due to iatrogenic trauma (10,60), that is, in about 2% of procedures. Of the patients in whom injury occurs, 17% have to be operated (60).

Among those undergoing a stenting procedure, 2 to 3% of uninfected patients will become septic (3,60,10). However, among patients with infected urinary tracts, approximately 31% will become septic (45,56,64). Preexistent urinary tract infection increases the risk of sepsis from urological instrumentation by 5 to 22 times over instrumentation in patients without prior infection (64).

After placement 1.5 to 3.8% of stents migrate or obstruct with debris (10,67). The likelihood of encrustation and obstruction increases with prolonged stenting, which occurs mostly with malignant disease (1).

We assume the hospital stay for retrograde ureteral stenting for acute urinary tract obstruction equals that for PCN, six days for an uncomplicated procedure and an extra six days for a complicated one.

c) Efficacy of RUS in sepsis:

We assume the efficacy of ureteral stenting in reducing mortality from urosepsis is 0.80, equal to that of a percutaneous nephrostomy tube, because the obstruction is relieved by either procedure.

d) Renal function:

As with percutaneous nephrostomy we assume that if drainage is performed immediately, and is successful, the probability of losing renal function is negligible.

5. Surgical intervention in pyonephrosis

Surgical intervention, either open surgical nephrostomy or nephrectomy, may be indicated if both PCN and RUS fail, or if serious ureteral injury caused by retrograde stenting occurs.

a) Operative mortality and efficacy of surgery in sepsis:

Death from nephrectomy and operative nephrostomy done for suppurative renal processes occurs in 6% of patients (18,21), and rises to 20 to 30% when done in the presence of a perinephric abscess (21). In the absence of pyonephrosis we assume the operative mortality is equal to that of cholecystectomy, that is 1.7% (41). The case fatality rate of gram-negative sepsis accompanying pyonephrosis, treated with surgical drainage is 12% (34), and thus the efficacy of surgery is 0.70.

b) Short term morbidity:

In patients without prior infection, complications occur in about 16% of operations, half of which are haemorrhage and injury (60). We assume the other half are infectious complications. On the other hand, among patients with preexistent infection 25% have complications (52), mostly exacerbation of infection. The average hospital stay for surgically treated patients is 13 days (34). If complications occur, we assume the hospital stay will be lengthened by another 13 days.

c) Nephrectomy / loss of renal function:

Among patients treated surgically for pyonephrosis, 75% undergo nephrectomy (23,70). In the remaining cases, open nephrostomy is performed. In patients undergoing open nephrostomy, loss of function again depends on the duration of obstruction before intervention.

6. Excess mortality rates

To calculate the utilities of the outcomes we need the excess mortality rate for calculus disease, for a solitary kidney and for a patient on renal dialysis. Patients with a history of urinary tract calculus disease have an excess mortality rate of 0.4/1000 per year (58). Patients with a solitary kidney have an excess mortality rate of 2.6/1000 per year (58). The National Dialysis Registry (58) provides excess mortality rates for renal dialysis depending on age and sex (see table 3).

7. Definitive treatment for a ureteral stone

Not all ureteral stones need to be treated. Depending on the stone's size and position it may pass spontaneously. If neither infection nor obstruction are present, most urologists will observe the patient for a few weeks. However, how long one may observe the patient without intervening is controversial, especially if signs of obstruction exist. Definitive treatment for a ureteral stone may be medical management, extracorporeal shockwave lithotripsy (ESWL), percutaneous nephrolithotomy (PCNL), ureteroscopic stone manipulation (URS) and ureterolithotomy.

The success rate of extracorporeal shockwave lithotripsy (ESWL) depends on the size of the stone, ranging from 53% for very large stones to 82% for small stones (12,16). 8% of procedures are complicated by obstruction, 1% by pyonephrosis and another 1% by sepsis, all of which are treated with percutaneous nephrostomy (16). The average hospital stay is two days, and another two days if complications occur.

Percutaneous nephrolithotomy (PCNL) can only be performed if the stone is situated in the renal pelvis or upper ureter. PCNL is very similar to percutaneous nephrostomy (PCN), except that in addition to cannulating the urinary tract, stone manipulation and retrieval takes place. Success is obtained in about 88% of cases (16,35). PCNL has a low mortality of 0.3% (37). As with PCN, sepsis occurs in 2.7% of procedures. Haemorrhage occurs more often than with PCN, namely in 1.2% (37). In addition, PCNL may be complicated by injury requiring surgery in 1.6% of procedures (35,37).

Ureteroscopic stone manipulation (URS) is similar to retrograde ureteral stenting (RUS). However, complications are slightly more frequent because in addition to passing the stone with a guidewire, stone crushing and stone retrieval instruments are used. Mortality occurs in 0.5%, sepsis in 3% and injury in 8% of cases (11,60). Success is obtained in 78%, that is, 51% of stones in the upper ureter and 86% of stones in the lower ureter are removed (16,60).

The mortality among patients undergoing elective ureterolithotomy lies between 0.1% (39) and 0.5% (41). Sepsis occurs in about 8% of operations and other short-term morbidity in another 8% (60). Ureterolithotomy fails in 1% of patients, in which case nephrectomy may be performed (43).

Table 2. Table of variables, baseline value and range of published values.

description	baseline value	range of published values	reference
cumulative probability of spontaneous passage of a stone, 1 to 1.5 years after onset of symptoms (see also figure 1):			
overall	0.73	0.64-0.80	55,9
upper ureter, < 4 mm	0.81		55
upper ureter, 4-6 mm	0.22		
upper ureter, > 6 mm	0		
lower ureter, < 4 mm	0.93	0.88-0.90	55,11
lower ureter, 4-6 mm	0.53	0.58-0.66	
lower ureter, > 6 mm	0.22	0.22-0.29	
rate of developing obstruction from a ureteral stone that does not pass, per day			
overall	0.0007		55
< 4 mm	0.0004		
4-6 mm	0.0006		
> 6 mm	0.0011		
rate of developing pyonephrosis given obstruction, per day			
overall	0.0010		55
< 4 mm	0.0005		
4-6 mm	0.0005		
> 6 mm	0.0019		
cumulative probability of developing sepsis given pyonephrosis	0.49	0.30-0.50 0.70	34,49 46
rate of developing sepsis given pyonephrosis, per day	1.573		53
probability of death due to sepsis, without intervention	0.385	0.15-0.51	13,32,34, 56,68

ULTRASOUND (US)			
dilatation on US, test for obstruction:			
true positive fraction	0.963		38,42,4,71
false positive fraction	0.260		15
likelihood ratio of a positive test	3.70		
likelihood ratio of a negative test	0.05		
internal echoes on US given a dilated system, test for pyonephrosis:			
true positive fraction	0.833		4,26,63
false positive fraction	0.060		
likelihood ratio of a positive test	13.88		
likelihood ratio of a negative test	0.18		
PREDICTIVE VALUES			
probability of obstruction at presentation:			
with minor symptoms	0.02		55
with flank pain/colic, haematuria	0.63		57
with dilatation on US	0.87		
without dilatation on US	0.08		
with stone, flank pain/colic, haematuria	0.85		57
with dilatation on US	0.95		
without dilatation on US	0.22		
with stone, severe flank pain/colic, anuria	1.00		
probability of pyonephrosis at presentation, in the presence of obstruction:			
with stone, colic, no fever	0.15		71
with echogenic material on US	0.71		
without echogenic material on US	0.03		
with stone, flank pain, fever, pyuria	0.86		71
with echogenic material on US	0.99		
without echogenic material on US	0.53		

PERCUTANEOUS NEPHROSTOMY (PCN)			
procedure related mortality	0.002	0	47
efficacy of PCN, mortality from sepsis prevented	0.80		34
probability of success of PCN			24,34,
in dilated tracts	0.96	0.92-0.99	42,52,61
in undilated tracts	0.88	0.85-0.91	38,42,52
second attempt after earlier failed PCN	0.33		24
post-PCN infection, overall	0.017		24,34, 42,52,61
post-PCN sepsis, overall	0.027	0.014	24,34, 42,52,61
post-PCN sepsis, in clinically suspected pyonephrosis	0.108		71
haemorrhage due to PCN	0.008	0.004-0.02	24,34, 42,52,61
increase in odds of haemorrhage and of death from PCN if bleeding diathesis exists	5		
loss of kidney if haemorrhage occurs (ie. embolization necessary)	0.20		24,34, 42,52,61
obstruction of tube / dislodgement / catheter breakage	0.02	0.006-0.017	24,34,61
average hospital stay, PCN	6 days		34
extra hospital stay for short-term morbidity	6 days		

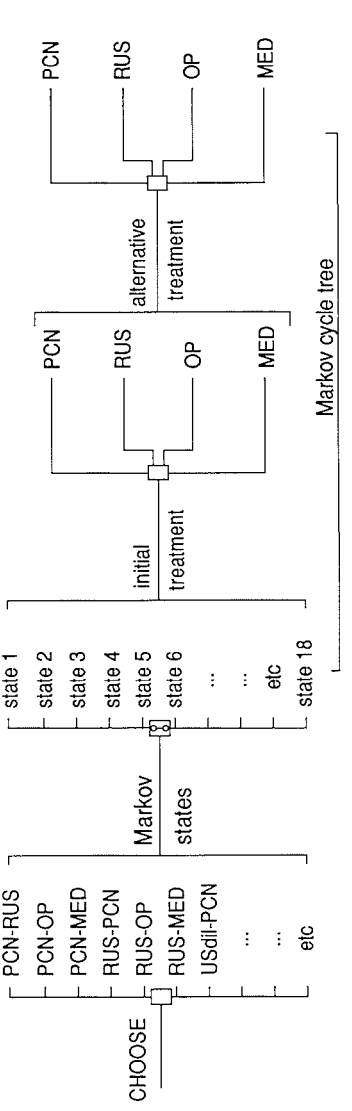
RETROGRADE URETERAL STENTING (RUS)			
procedure related mortality	0.005	0	60,11
efficacy of RUS, assumed equal to efficacy of PCN	0.80		
probability of success of RUS	0.94	0.80	11,40
post-RUS sepsis, overall	0.02		3,10,60
post-RUS sepsis, in clinically suspected pyonephrosis	0.101	0.308	64
iatrogenic trauma	0.02		3,10,60
operation necessary if iatrogenic trauma occurs	0.17		60
migration / obstruction of stent	0.015	0.038-0.135	67,1,10
average hospital stay, RUS (same as for PCN)	6 days		
extra hospital stay for short-term morbidity (same as for PCN)	6 days		

SURGERY (OPER)			
operative mortality of a comparative abdominal procedure	0.017		41
suppurative renal process	0.06	0.03-0.30	18,21,2
efficacy of surgery, decrease in mortality due to sepsis	0.70		34
short-term morbidity,			
non-infected tract	0.16		60
infected tract	0.25		52
probability of sepsis due to surgery			
non-infected tract	0.08		52,60
infected tract	0.17		
average hospital stay, operation, without short-term morbidity	13 days		2,34
extra hospital stay for short-term morbidity	13 days		
nephrectomy is necessary/done as opposed to surgical nephrostomy	0.62	0.75	70,23

Table 3. Excess mortality rates per 1000 per year

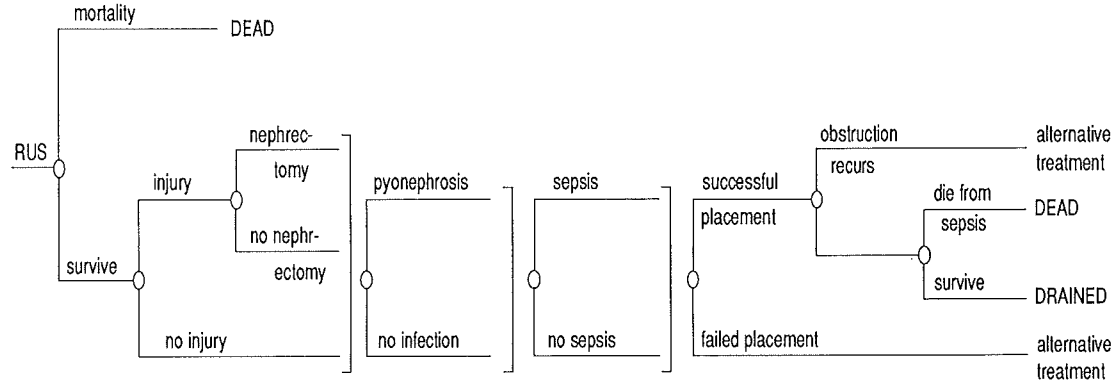
for calculus disease		0.0004		58
for solitary kidney		0.0026		58
for renal dialysis,	AGE	MALE	FEMALE	58
	15	0.083	0.094	
	25	0.087	0.094	
	35	0.115	0.114	
	45	0.125	0.114	
	55	0.138	0.130	
	65	0.157	0.169	

Figure 4.

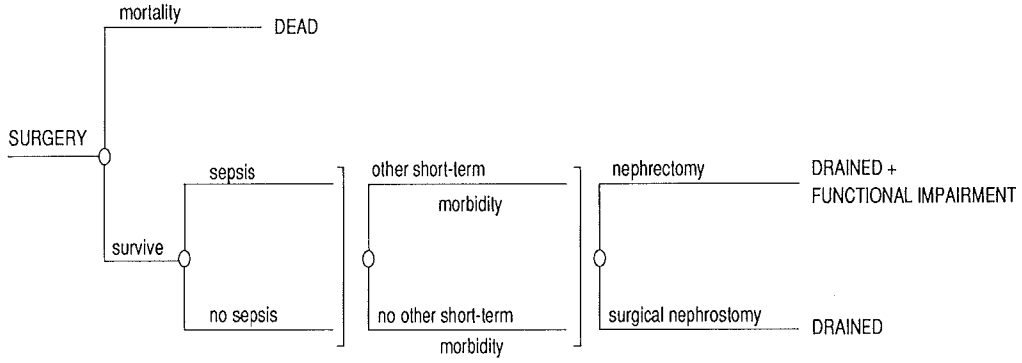


1. The decision model: a) overview of the model. For abbreviations of the strategies see table 1.
□ represents a choice node ⚅ represents a Markov node ○ represents a chance node

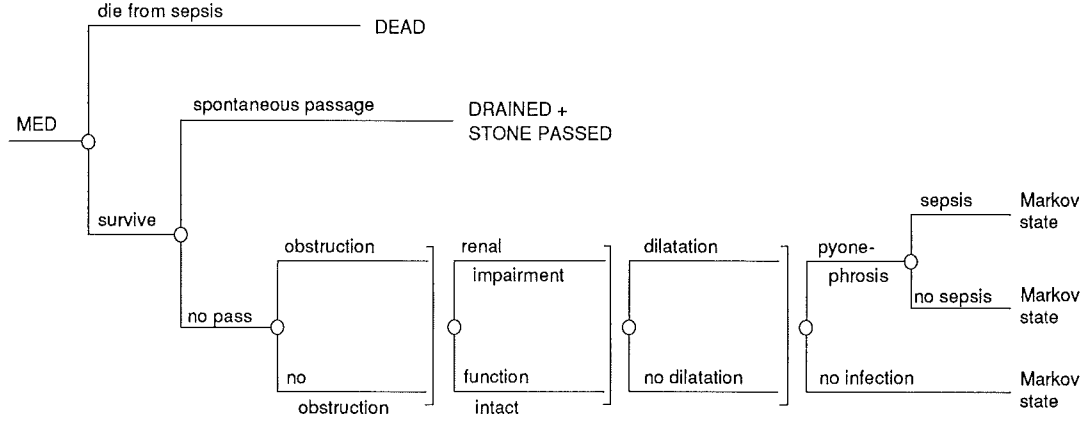
1b) Subtree for percutaneous nephrostomy (PCN)



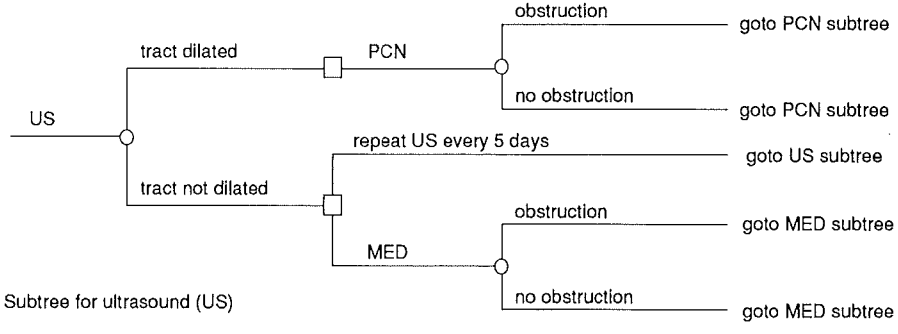
1c) Subtree for retrograde ureteral stenting (RUS)



1d) Subtree for surgical management.



1e) Subtree for medical management.



1f) Subtree for ultrasound (US)

Table 4. Table of the Markov states of the model.

state	description		
1	no obstruction	urolithiasis	renal function intact
2	obstruction	urolithiasis, dilated	renal function intact
3	obstruction	urolithiasis, dilated	loss of function
4	obstruction	urolithiasis, non-dilated	renal function intact
5	obstruction	urolithiasis, non-dilated	loss of function
6	pyonephrosis	urolithiasis, dilated	renal function intact
7	pyonephrosis	urolithiasis, dilated	loss of function
8	pyonephrosis	urolithiasis, non-dilated	renal function intact
9	pyonephrosis	urolithiasis, non-dilated	loss of function
10	sepsis	urolithiasis, dilated	renal function intact
11	sepsis	urolithiasis, dilated	loss of function
12	sepsis	urolithiasis, non-dilated	renal function intact
13	sepsis	urolithiasis, non-dilated	loss of function
14	drained	urolithiasis	renal function intact
15	drained	urolithiasis	loss of function
16	drained	stone passed	renal function intact
17	drained	stone passed	loss of function
18	dead		

IV. THE MODEL

The decision model combines a Markov process (7) with a recursive decision tree. The time horizon for the Markov process is 50 days. The remaining prognosis (the 'tail utility') is calculated with the DEALE-method (5,6). A Markov process simulates a cohort of patients as they move from one state of health to another. Time is divided into slices called Markov cycles. Table 4 summarizes the 18 Markov states of health for acute urinary tract obstruction caused by urolithiasis. The states of health in this analysis can be characterized by three features and their complications: 1) presence or absence of obstruction and its complications, pyonephrosis and gram-negative sepsis; 2) presence or absence of urolithiasis, with or without dilatation of the tract; and 3) function of the affected kidney. If the tract is successfully drained by means of an interventional procedure, or the stone passes spontaneously, obstruction and infection are assumed to be relieved.

Figure 4a gives an overview of the decision model and figures 4b through f show the subtrees for PCN, RUS, surgery, medical management and ultrasound respectively. The model starts with a decision node (represented with a square) on the left with 18 strategies (see table 1). The Markov process is represented by the Markov node (a square with two circles), the 18 states that characterize the states of health in this analysis and the Markov cycle tree. The cycle tree starts with the initial treatment and branches according to the ensuing events. An alternative treatment may follow depending on whether the procedure is successful, whether obstruction recurs, or whether complications develop during medical management.

The subtrees for PCN and RUS (figure 4b and c) are very similar. The subtrees start on the left with a chance node (represented with a circle) for mortality and branch successively depicting the events that may occur. The procedure may result in mortality or morbidity. Major morbidity includes haemorrhage following PCN and iatrogenic injury following RUS. If life threatening haemorrhage occurs following PCN, the renal artery is embolized, in which case function of the kidney is lost. If serious iatrogenic injury occurs following RUS, surgical intervention may be necessary. The procedure may cause infection or exacerbate an existing infection. Placement of the drainage tube may be successful or not. Even if placement is initially successful, the tube may obstruct or break, in which case urinary tract obstruction will recur. If tube placement is successful, death from sepsis is prevented in a proportion of the patients. If placement is unsuccessful or obstruction recurs, an alternative treatment may be chosen.

The surgery subtree (figure 4d) starts on the left with a chance node depicting operative mortality. The following branches represent the possible development of sepsis and/or other short-term morbidity, and whether surgical intervention consists of nephrectomy or nephrostomy.

As with the other subtrees, the events which may occur during medical management are represented with chance nodes (circles) and successive branching (figure 4e). Medical management may result in death from sepsis, depicted with a chance node on the left. In a proportion of cases, however, spontaneous passage of the stone occurs, in which case the obstruction and infection are relieved. If obstruction is not present initially, it may develop in due course as a result of the stone. Depending on the duration of obstruction, renal function may become impaired. A non-dilated tract may dilate during expectant management and pyonephrosis and sepsis can complicate the course of a non-invasive approach. Depending on the initial circumstances and on the chosen strategy (table 1), if sepsis or pyonephrosis occur, or if the stone does not pass within 5 to 15 days, a drainage procedure is performed.

Figure 4f presents the diagnostic strategy "if an ultrasound shows a dilated tract, perform PCN". The subtree starts on the left with a chance node for the test result: either ultrasound shows dilatation or not. If the ultrasound shows dilatation, PCN is performed. Patients with non-dilated tracts are managed medically, with or without repeating the ultrasound periodically. Although a dilated tract is usually a sign of obstruction, non-obstructed tracts may be dilated, in which case PCN will be performed although obstruction does not exist. In addition, some obstructed tracts are not dilated, in which case intervention is postponed.

With some small modifications, the presented model applies to the decision whether or not a patient with a ureteral stone should undergo definitive treatment. Treatment options are medical management, extracorporeal shockwave lithotripsy (ESWL), percutaneous nephrolithotomy (PCNL), ureteroscopic stone manipulation (URS) and ureterolithotomy. The ESWL subtree is analogous to that of medical management. Obstruction, pyonephrosis and sepsis may complicate ESWL, in which case PCN will be performed. Percutaneous nephrolithotomy (PCNL) can only be performed if the stone is in the renal pelvis or upper ureter. The procedure is very similar to PCN. In addition to the complications of PCN, PCNL may be complicated by injury requiring surgery. Ureteroscopic stone manipulation (URS) is very similar to retrograde ureteral stenting (RUS). Utilizing the analogous structure of the choice of temporary drainage for obstruction and the choice of definitive treatment for urolithiasis, we also examined the latter decision.

V. TECHNICAL DETAILS

The probabilities modelled above and presented in graphs 1 to 4, are cumulative probabilities, or cumulative incidence functions, of the event. To perform the calculations of the Markov process we need to know the transition probability, in other words the probability that the event will occur during one cycle. To calculate the transition probability, the cumulative probability function is converted to the corresponding hazard rate h(t) (or incidence density function) with the following equation (36):

$$h(t) = \frac{F'(t)}{1 - F(t)}$$

Subsequently, the transition probability p(i) during cycle i, from time t_{i-1} to t_i, is calculated as follows:

$$p(i) = 1 - e^{-\int_{t_{i-1}}^{t_i} h(t)dt}$$

If event B is conditional on prior event or state A, then the transition probability for event B as calculated with the above equation, applies for all patients in state A since the beginning of the Markov process. However, if event or state A is initially absent but occurs subsequently during cycle x, the hazard rate for event B during cycle i, is $h(t_{i-x})$. For example, the probability of renal impairment depends on the duration of obstruction and, thus, if obstruction is not present initially but develops during cycle x, the probability of renal impairment will depend on the period of time elapsed since cycle x. Furthermore, obstruction may have developed during any cycle x from 0 to i. Thus, the group of patients with obstruction is a heterogeneous mixture of patients with obstruction of varying duration, each with a corresponding probability of renal impairment. The transition probability of renal impairment (event B) during cycle i is, therefore, a summation

of products. The products are the probability of obstruction (event A) during cycle x, multiplied by the probability of renal impairment (event B) given obstruction during i-x cycles. The products are summed over all cycles x from 0 to i, in equation:

$$pB(i) = \sum_{x=0}^{i} pA(x) \cdot pB \mid A(i-x)$$

As discussed in section III.1.d) the rate of obstruction, and therefore the transition probability during each cycle, is constant. Therefore, if p(obstruction) is the transition probability of obstruction during a cycle, then:

$$pB(i) = p(obstruction) \cdot \sum_{x=0}^{i} pB \mid obstruction(i-x)$$

This equation applies to the transition probability of renal impairment and of dilatation, because both depend on the duration of obstruction. By calculating the transition probability to dilatation and to renal impairment in this way, we actually introduce a form of memory into our model, making it a semi-Markov process rather than a Markov process. The method is also known as a Markov memory matrix (59). An alternative method would be to create a new state for every duration of obstruction, which would, however, increase the number of states to 126, making the model impractical.

VI. ASSUMPTIONS

To make the problem tractable, we make the following simplifying assumptions:

1. Pyonephrosis implies an infected obstructed renal collecting system.
2. The pathophysiology of hydronephrosis in humans is similar to that in dogs.
3. A kidney functioning at less than 20% of normal is assumed to be a remnant kidney.
4. The probability of losing renal function with immediate successful drainage is negligible. If intervention is postponed, the probability of functional loss increases with the duration of obstruction. If obstruction recurs following a drainage procedure, the probability of renal impairment depends on the period of time elapsed since the initial obstruction.
5. If the stone passes spontaneously, or if the tract is successfully drained by means of an interventional procedure, obstruction and infection are relieved.
6. If the stone neither passes spontaneously nor is removed, and the tract is not drained, obstruction and infection will eventually occur.
7. The efficacy of ureteral stents in reducing mortality from urosepsis equals that of a percutaneous nephrostomy tube.
8. If the patient develops sepsis as a complication of a procedure, the drainage procedure helps resolve the sepsis with the same efficacy as for patients with sepsis prior to the procedure.

9. All false positive ultrasound results for the diagnosis obstruction will be false positive at presentation and, therefore, dilatation on ultrasound that develops during the time horizon of the analysis implies obstruction of the tract.

10. Definitive treatment will be instituted within 7 weeks.

VII. RESULTS

1. The best strategies

Using the baseline values in table 2, and folding back and averaging out for each of the examples, we calculate the following results.

Case 1: For a 40 year old male presenting with a small stone in the lower ureter, clinical signs of pyonephrosis, dilatation on US but no echogenic material in the dilated tract, the best strategy is non-operative intervention, with either PCN or RUS (table 5). PCN is preferred over RUS by only 0.09 life years (0.3% of the outcome value). Either procedure is better than surgery by at least 2.5 years and better than medical management by at least 7.8 years. Postponing intervention until sepsis occurs yields 1.3 years less. If PCN is done initially but subsequently obstruction recurs, or if placement of the tube fails, RUS should be performed.

Case 2: For a 65 year old male with urosepsis presumably due to an obstructing stone and an increased bleeding tendency, our model suggests that RUS should be performed (table 5). RUS is safer than PCN for this patient, because the risks of haemorrhage are greater for PCN than RUS. However, if RUS is not available, PCN should be performed in spite of the increased bleeding tendency. The difference in life expectancy for these two strategies is only 0.06 years (0.5% of the outcome). Both procedures are better than surgical management by at least 0.6 years and better than medical management by at least 8 years.

Case 3: Our model suggests that a 60 year old male with a staghorn calculus in a solitary kidney, who presents with anuria, flank pain and urosepsis after ESWL, is best treated with PCN or RUS (table 5). The difference in life expectancy between these two strategies is 0.04 years (0.3% of the outcome). Surgical and medical management yield, respectively, 6.5 and 14 fewer life years.

Case 4: A 40 year old man with a small stone in the lower ureter, renal colic and dilatation on US, is best managed medically (table 5). If pyonephrosis or sepsis occur, or if the stone has not passed after 15 days, intervention is indicated. However, the difference in expected utility with immediate drainage is only 0.06 life years (0.2% of the outcome). Medical management without intervention yields 1.4 fewer life years and surgery yields 2 years less. Postponing drainage until signs of pyonephrosis occur yields 0.9 fewer years.

Table 5. Expected utility, that is life expectancy in years, of the most relevant strategies for the four case examples. For abbreviations of the strategies see table 1.

CASE	STRATEGY	LIFE EXPECTANCY
1: 40 year old male,	PCN-RUS	32.44
4 mm stone in the lower ureter,	RUS-PCN	32.35
clinically pyonephrosis,	MED-sep-PCN	31.15
dilated tract	OPER	29.80
	MED	24.48
2: 65 year old male,	RUS-PCN	12.76
flank pain, urosepsis,	PCN-RUS	12.70
increased bleeding tendency,	OPER	12.09
mildly dilated tract,	MED	4.51
3: 60 year old male,	PCN-RUS	14.76
staghorn calculus, post ESWL,	RUS-PCN	14.72
anuria, flank pain, urosepsis,	OPER	8.14
solitary kidney	MED	0.60
4: 40 year old male,	MED-time-PCN	32.60
4 mm stone in the lower ureter, renal colic,	PCN-RUS	32.54
haematuria,	MED-pyo-PCN	31.73
dilated tract	MED	31.18
	OPER	30.60

A number of the values used in the analysis are uncertain. Sensitivity analysis is the process of calculating the expected utility of the strategies for a range of values of a variable to determine if altering the value affects the choice of optimal strategy. We analyzed the effect of different values of the probabilities using this method.

2. Probability of pyonephrosis and obstruction at presentation

Figure 5a presents the results of a one-way sensitivity analysis for the first case example for the probability of pyonephrosis at presentation, given obstruction. The x-axis represents the variable analyzed, ie. the probability of pyonephrosis, and the y-axis represents the expected utility expressed as life expectancy in years. Each line represents the expected utility for a particular strategy. As the probability of pyonephrosis increases, the expected utility of each strategy falls. In addition, the higher the probability of pyonephrosis, the larger the difference between an

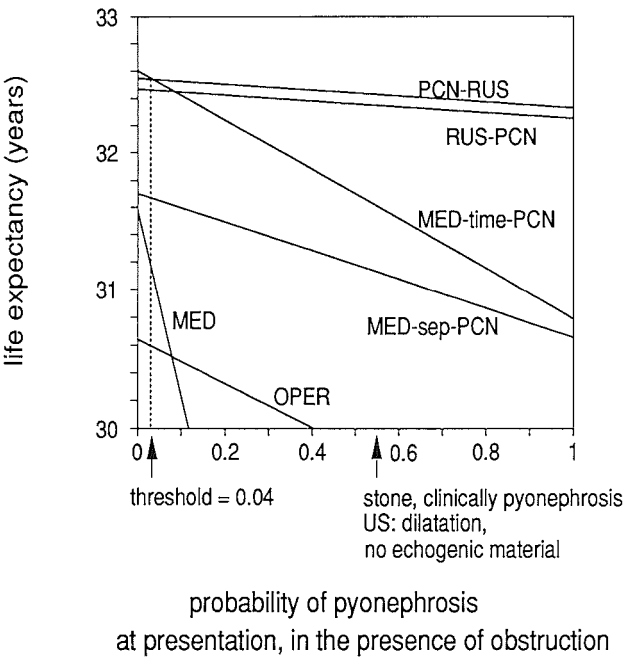


Figure 5a. One-way sensitivity analysis for the probability of pyonephrosis at presentation in the presence of obstruction, for a 40 year old man with a small stone in the lower ureter, dilatation of the collecting system and clinical signs of pyonephrosis (case 1). For abbreviations of the strategies see table 1.

interventional strategy and medical management, thus increasing the benefit of early intervention. A patient with signs of pyonephrosis but no echogenic material visible in the collecting system on ultrasound, has a probability of pyonephrosis of 0.53 (table 2), indicating (figure 5a) intervention as preferred strategy. If, however, pyonephrosis is doubtful, with a probability of lower than 0.04, observing the patient for 15 days seems worthwhile, giving the stone a chance to pass spontaneously.

Figure 5b presents the results of a one-way sensitivity analysis for the probability of obstruction for the fourth case. The probability of obstruction may be estimated based on clinical signs and symptoms, the duration of symptoms, and the result of an ultrasound examination. For example, a patient with a ureteral stone, renal colic, haematuria and dilatation on ultrasound has a probability of obstruction of 0.95 (table 2), and therefore (figure 5b) the patient may be treated

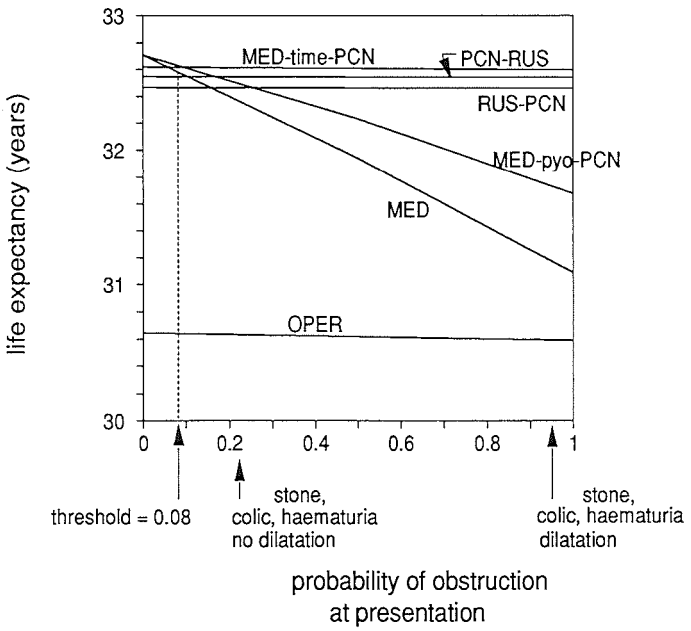


Figure 5b. One-way sensitivity analysis for the probability of obstruction at presentation for a 40 year old man with a small stone in the lower ureter, with dilatation of the collecting system on ultrasound but without clinical signs of pyonephrosis (case 4). For abbreviations of the strategies see table 1.

medically for 15 days; however, intervention is necessary if pyonephrosis or sepsis occur, or if the stone does not pass. If the probability of obstruction is below 0.08, the patient may be treated medically with intervention being necessary only if complications occur.

Figure 5c presents a two-way sensitivity analysis for the probability of obstruction and of pyonephrosis. With a two-way sensitivity analysis two parameters are varied simultaneously. The graph defines areas where, depending on the combination of the probabilities, a strategy is optimal. If both the probability of obstruction and pyonephrosis are low (the left bottom corner of the graph marked MED-time-PCN) one should postpone intervention and observe the patient, to give the stone a chance to pass spontaneously. However, for higher probabilities (the area marked PCN-RUS), one should intervene immediately.

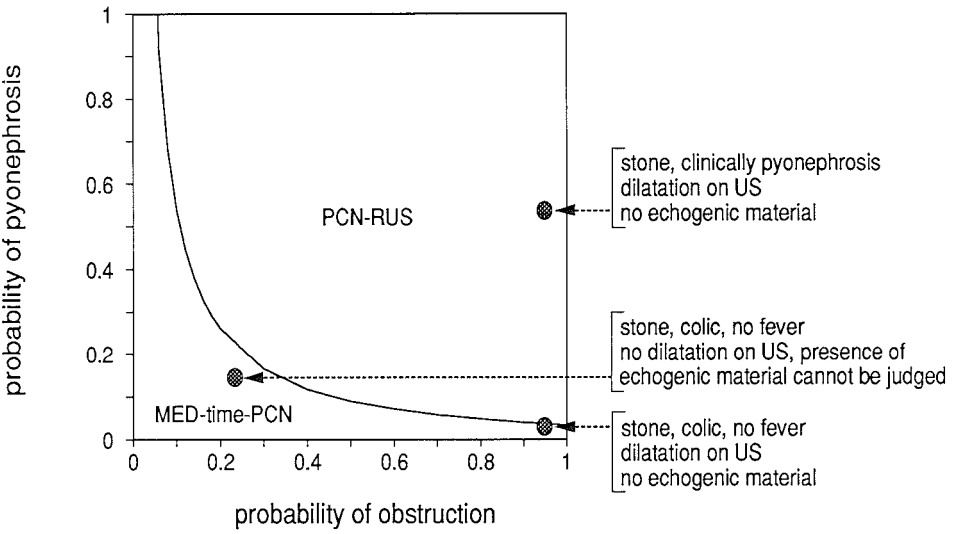


Figure 5c. Two-way sensitivity analysis for the probability of obstruction and of pyonephrosis at presentation, for a 40 year old man with a small stone in the lower ureter. For abbreviations of the strategies see table 1.

3. Ultrasound examination and dilatation of the collecting system

In clinical practice the decision to intervene is often based on the presence of dilatation on ultrasound (US) examination. We analyzed this approach by including the strategies 1) if an ultrasound shows dilatation perform PCN, otherwise manage medically and 2) if an ultrasound shows dilatation perform PCN, otherwise manage medically and repeat the ultrasound periodically with subsequent PCN if dilatation occurs. Whether or not one should perform an ultrasound examination depends on the prior probability of obstruction (figure 6). Patients with a prior probability of obstruction lower than 0.02 should be managed medically, regardless of the ultrasound result. For prior probabilities above 0.02 an ultrasound should be done with a positive result dictating intervention, and a negative result medical management and observation with periodic ultrasounds. Above a prior probability of 0.58, drainage should be performed if the stone has not passed within 15 days, irrespective of the US result. If observation and repeating the ultrasound is impractical, the latter threshold shifts to 0.42.

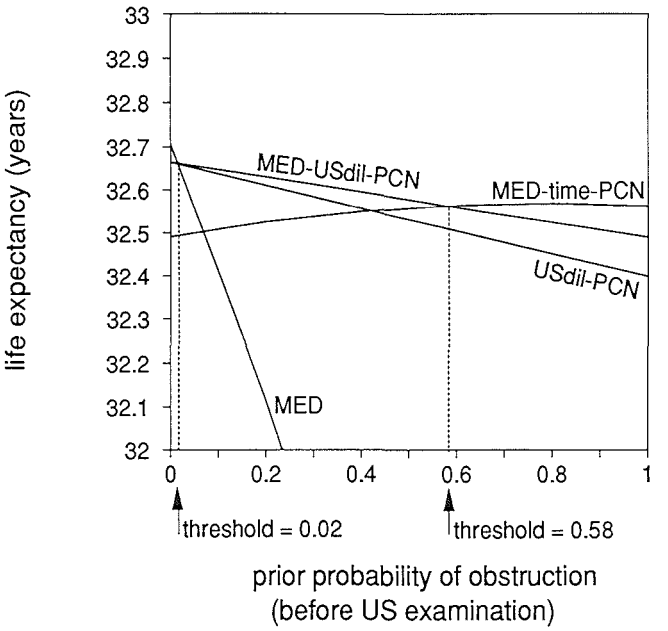


Figure 6. One way sensitivity analysis for the prior probability of obstruction for the strategies 1) if ultrasound reveals dilatation perform PCN (USdil-PCN) 2) medical management with ultrasound examination every 5 days, if ultrasound reveals dilatation, then perform PCN (MED-USdil-PCN) 3) medical management (MED) and 4) medical management, PCN if pyonephrosis or sepsis occur or if the stone has not passed after 15 days (MED-time-PCN).

Our model suggests that if pyonephrosis is likely, postponing PCN until dilatation develops, to increase the chance of successful placement of a nephrostomy tube, does not provide any benefit over immediate PCN. The probability of successfully placing a PCN tube in an undilated tract is only slightly lower than in a dilated tract (88% compared to 96%). Therefore, PCN may be attempted even though dilatation is not present. The slight decrease in successful placement in the absence of dilatation is outweighed by the risk of sepsis and renal failure.

4. Mortality due to sepsis and the probability of developing sepsis

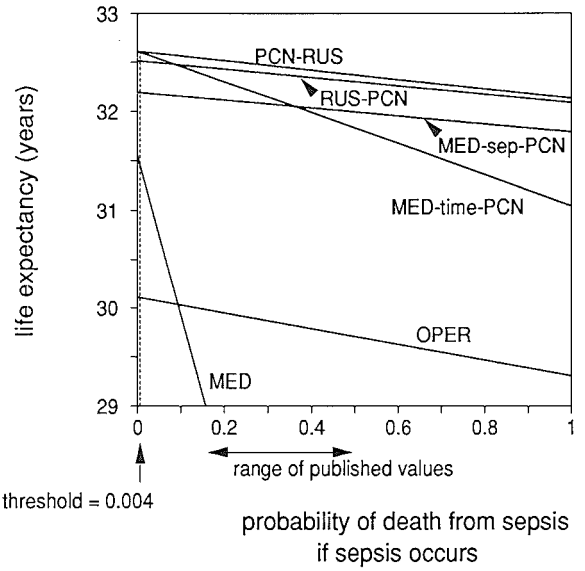


Figure 7a. One-way sensitivity analysis for the probability of death if gram-negative sepsis were to occur, for a 40 year old man with a small stone in the lower ureter, dilatation of the collecting system and clinical signs of pyonephrosis (case 1).

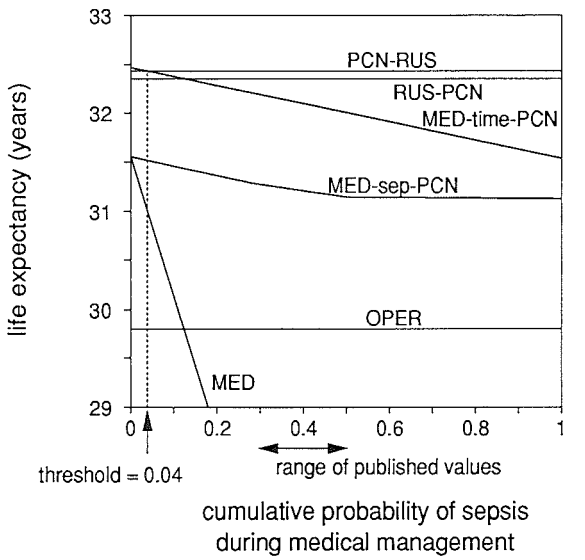


Figure 7b. One-way sensitivity analysis for the cumulative probability of sepsis during medical management, for a 40 year old man with a small stone in the lower ureter, dilatation of the collecting system and clinical signs of pyonephrosis (case 1).

If the clinical presentation suggests pyonephrosis (eg. case 1) and the mortality due to sepsis is greater than 0.004 (figure 7a), or the probability of developing sepsis is greater than 0.04 (figure 7b), intervention provides a higher survival. The ranges of published values, for both the mortality due to sepsis and the probability of developing sepsis, lie far above the calculated thresholds, implying that the decision is not sensitive to the precise value of these probabilities.

5. Solitary kidney

For patients with an obstructed infected solitary kidney, intervention is indicated to prevent renal failure from prolonged obstruction. For example, case 3 presents a patient with a solitary kidney, for whom intervention is better than an expectant approach. For cases 1 and 2, assuming they have a solitary kidney, intervention is also preferred to an expectant approach. Furthermore, the difference in life expectancy between intervention and medical management increases (10 to 50%) if the patient has a solitary kidney as opposed to two kidneys. Renal function of a solitary obstructed kidney remains intact longer than that of one of a pair, presumably due to a compensatory mechanism. However, losing function of a solitary kidney has graver consequences. The loss of a solitary kidney results in renal failure and necessitates dialysis. Therefore, pyonephrosis in a solitary kidney calls for prompt intervention. However, our model suggests that patients with an obstructed non-infected solitary kidney may be managed medically, intervention being necessary should complications occur.

6. Increased bleeding tendency and increased risk of PCN/RUS

Urosepsis is frequently associated with an increased bleeding tendency. For patients with an increased bleeding tendency, RUS has a higher expected utility than PCN (table 6), because PCN is more likely to be complicated by haemorrhage which may require embolization of the renal artery. If the odds of haemorrhage are increased by a factor of five, the difference in survival between PCN and RUS is approximately 0.2 years (that is 0.5% of the outcome value). In general, if a patient has an increased bleeding tendency, RUS should be attempted first, if this fails PCN may be performed. The potential benefit of PCN outweighs the risks of medical or surgical management, even with an increased risk of haemorrhage.

Table 6. Results of the analysis depending on whether there is dilatation and signs of pyonephrosis or sepsis for a 40 year old male with a stone in the lower ureter and an increased bleeding diathesis. For the abbreviations of the strategies see table 1.

SIGNS and SYMPTOMS	STRATEGY	LIFE EXPECTANCY
colic, haematuria, no dilatation on US, no fever	MED-time-RUS MED-pyo-RUS RUS-PCN PCN-RUS	32.47 32.44 32.43 32.24
colic, haematuria, dilatation on US, no fever	MED-time-RUS RUS-PCN PCN-RUS MED-pyo-RUS	32.44 32.42 32.24 31.22
clinically pyonephrosis, dilatation on US, no echogenic material	RUS-PCN PCN-RUS OPER MED	32.33 32.16 29.80 21.18
urosepsis	RUS-PCN PCN-RUS OPER MED	30.35 30.19 28.00 11.07

In the model we linked the risk of death from PCN to the risk of haemorrhage so that increasing the odds of death increases the odds of haemorrhage by the same proportion. An increase in the mortality of PCN is most likely to be due to an increase in the risk of haemorrhage. By linking these two probabilities we are able to perform a sensitivity analysis on both risks simultaneously by varying the probability of death. The threshold value of PCN mortality above which RUS is preferred is 0.005. Similarly, we linked the risk of death from RUS to the risk of injury and performed a sensitivity analysis on the mortality of RUS. The threshold value of RUS mortality above which PCN is preferred is 0.002. Both PCN and RUS are low risk procedures and the benefit of drainage clearly outweighs the low risk of these procedures when faced with pyonephrosis. Figure 8 presents a two-way sensitivity analysis showing that the procedure related mortality must be quite high before medical management is preferred.

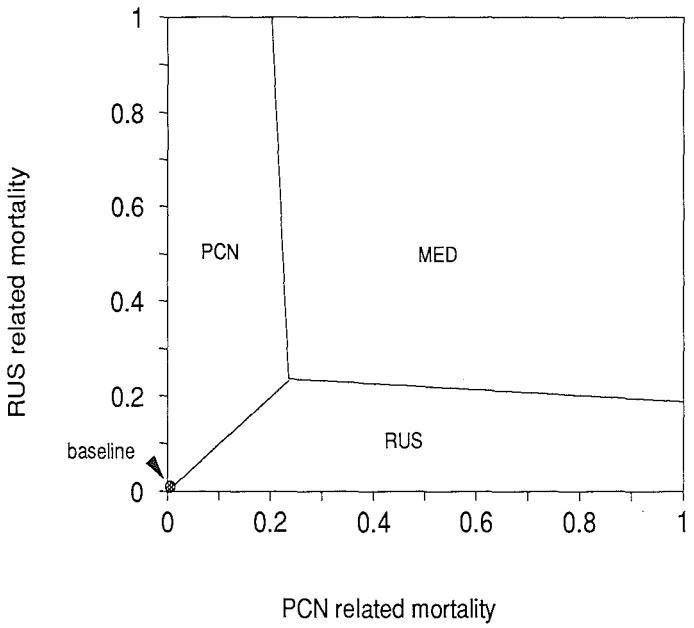


Figure 8. Two-way sensitivity analysis for the probability of death from PCN and the probability of death from RUS, for a 40 year old man with a small stone in the lower ureter, dilatation of the collecting system and clinical signs of pyonephrosis (case 1).

7. Size and position of the stone

Table 7 summarizes the results of the analysis for different stone sizes and positions, for patients with a normal bleeding tendency. Patients with or without dilatation, patients with clinical signs of pyonephrosis and patients with sepsis are considered. Patients without dilatation of the tract and without signs of pyonephrosis, should be managed medically, a drainage procedure being indicated should signs of pyonephrosis or sepsis develop, unless the stone is large and in the upper ureter. Large stones in the upper ureter virtually never pass spontaneously and, thus, intervention is indicated even in the absence of dilatation. Dilatation on ultrasound, caused by a large stone in the upper or lower ureter calls for intervention. However, patients with dilatation caused by a small or medium sized stone in the upper or lower ureter, may be observed for 15 days, giving the stone a chance to pass spontaneously. Obstruction with pyonephrosis or sepsis, whatever the size or position of the stone, necessitates intervention to prevent death from sepsis and prevent functional impairment of the affected kidney.

Table 7. Results of the analysis depending on position and size of the stone, whether there is dilatation, clinical signs of pyonephrosis or sepsis for patients with a normal bleeding diathesis. For the abbreviations of the strategies see table 1.

position and size of the stone	colic, haematuria, no fever, no dilatation	colic, haematuria, no fever, dilatation, no echogenic material	clinically pyonephrosis, dilatation, no echogenic material	urosepsis[3]
upper ureter				
< 4 mm	MED-time-PCN	MED-time-PCN	PCN-RUS	PCN-RUS
4-6 mm	MED-time-PCN	MED-time-PCN	PCN-RUS	PCN-RUS
> 6 mm	PCN-RUS	PCN-RUS	PCN-RUS	PCN-RUS
lower ureter				
< 4 mm	MED-time-PCN[4]	MED-time-PCN	PCN-RUS[1]	PCN-RUS
4-6 mm	MED-time-PCN	MED-time-PCN	PCN-RUS	PCN-RUS
> 6 mm	MED-time-PCN	PCN-RUS	PCN-RUS	PCN-RUS

[1] case 1
[3] case 3
[4] case 4

8. Definitive treatment and timing of intervention

As explained, the models for a temporary drainage procedure and for definitive treatment are very similar. Modifying the model to examine definitive treatment and analyzing the problem for case 4 at the time of onset of symptoms, we find that the patient should be treated medically. Depending on the size and position of the stone, patients may be treated in various ways (table 8). For a given size and position of the stone, the reported strategies differ little in expected utility. However, the results depend on the procedure related mortality. An extensive sensitivity analysis of the relevant probabilities of definitive treatment is beyond the scope of this article.

Table 8. Results of the analysis, examining definitive treatment, depending on position and size of the stone and whether ultrasound reveals dilatation. The treatment options are: medical management, ultrasound every 15 days, if dilatation occurs treat (MED-dil-TREAT); medical management, if the stone has not passed after 15 days treat (MED-TREAT); extracorporeal shockwave lithotripsy (ESWL); percutaneous nephrolithotomy (PCNL); ureteroscopic stone removal (URS) or ureterolithotomy (LITH). Treatment strategies consist of an initial treatment and an alternative, if the initial fails. The strategies are tabulated in order of decreasing expected utility. Only the strategies with an expected utility close to the best strategy are reported, that is with a difference in life expectancy of less than one week.

position and size of the stone	no dilatation on ultrasound	dilatation on ultrasound
UPPER URETER		
< 4 mm	MED-dil-ESWL	MED-LITH
		MED-ESWL
		MED-PCNL
4-6 mm	MED-dil-ESWL	ESWL-LITH
		ESWL-PCNL
		PCNL-LITH
		MED-LITH
		MED-ESWL
		MED-PCNL
> 6 mm	MED-dil-ESWL	ESWL-LITH
	ESWL-LITH	ESWL-PCNL
		PCNL-LITH
LOWER URETER		
< 4 mm	MED-dil-ESWL	[4]MED-LITH
		MED-ESWL
4-6 mm	MED-dil-ESWL	MED-LITH
		MED-ESWL
		ESWL-LITH
> 6 mm	MED-dil-ESWL	ESWL-LITH
		ESWL-URS

[4] case 4

However, how long one may postpone intervention does relate to the issue of this paper. Figure 9 shows a sensitivity analysis for the time one postpones intervention for a 40 year old man with colic, haematuria and dilatation on ultrasound. The horizontal axis represents the number of days intervention is postponed. The curves plot life expectancy as a function of the number of

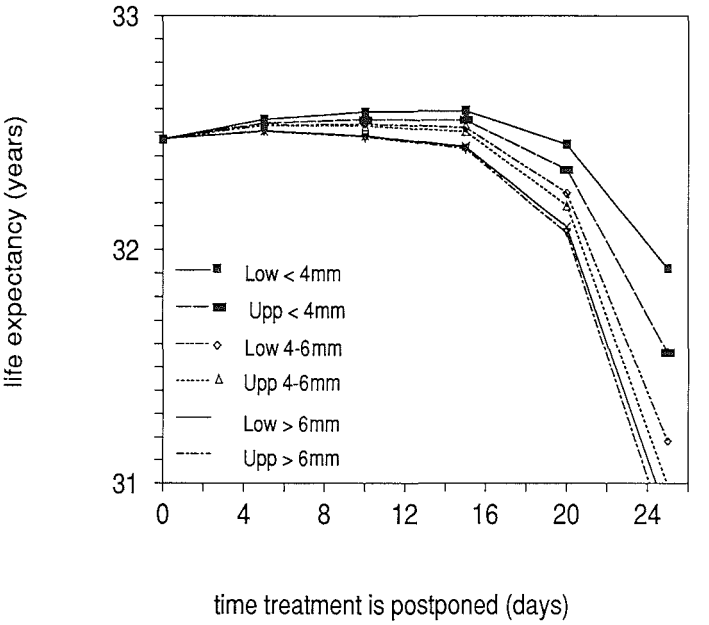


Figure 9. Sensitivity analysis for the time one postpones treatment for a patient with dilatation of the collecting system, without clinical signs of pyonephrosis, for small (<4 mm), medium (4-6 mm) and large (>6 mm) stones in the upper (Upp) and lower (Low) ureter.

days definitive treatment is postponed. The curves shown are for small, medium and large stones in the lower and upper ureter. The length of time one may wait to allow for spontaneous passage varies from 15 days for small stones in the lower ureter to 5 days for large stones in the upper ureter. If the stone has not passed within 5 to 15 days, the likelihood that it will pass spontaneously is very small and the probability of complications increases, especially if the stone is large and in the upper ureter. The strategy of medical management followed by a drainage procedure follows a similar curve as for medical management followed by definitive treatment, except for a slightly lower expected utility. The reason for a lower expected utility is that a proportion of patients will not pass the stone spontaneously after drainage and will still have to undergo another procedure to remove the stone, implying that the drainage procedure incurs risk without the benefit of definitive treatment.

VIII. DISCUSSION

Using a decision analytic model, we have examined whether to intervene or adopt an expectant approach in patients with acute urinary tract obstruction. Many of the relevant probabilities, such as the probability of loss of renal function, the probability of dilatation and the probability that the stone will pass spontaneously are time-dependent, either related to the duration of obstruction or to the time since onset of symptoms. Because of the time dependency of the probabilities, we used a Markov process. Markov processes are conventionally used to model the prognosis of patients with chronic diseases. Clearly, the problem addressed in this paper is acute rather than chronic. Even so, we found using a Markov process to model part of the problem convenient for this analysis, because it allows for modelling of a process with repetitive risks that evolve over time.

Our model is limited by several difficulties inherent in deriving probability estimates from the literature. The medical literature often reports success and complication rates of new techniques as overall experience or experience after the learning phase has passed. Which estimate then is the appropriate one to use? The overall experience is less optimistic, but probably reflects the average hospital more reliably. After all, the published reports are usually from clinics with more experience in the technique than that of the average hospital. Another problem with finding accurate data is that entities are often defined differently by the different authors. Some define pyonephrosis, for example, as an infected obstructed renal collecting system; others only apply this term to grossly affected and non-functioning kidneys. Moreover, if authors do not clearly state the definitions used, readers cannot be certain which patient groups their results are applicable to. The denominator problem is a similar issue. Studies often do not report explicitly which cohort the denominator refers to when calculating percentages. This can be misleading and lead to inappropriate relative frequencies.

A related problem is that of double counting. For example, a death may be counted as a PCN death, and at the same time as an operative death. Another example of double counting occurs when authors use the same cases for multiple publications.

We based our estimates of the effect of obstruction duration to dilatation of the urinary tract on animal experiments. We assumed that the pathophysiology of hydronephrosis in dogs is similar to that in humans. Is this assumption valid? And are the derived estimates appropriate? Because clinical judgement is partly based on knowledge from these animal experiments, and because it is the closest approximation to human physiology that is available, we used data based on these experiments.

Apart from the limitations of our model inherent in deriving probability estimates from the literature, our model is restricted to the problem of acute urinary tract obstruction caused by urolithiasis. We have chosen not to analyze in depth the associated clinical decision concerning the definitive treatment of the ureteral stone. We briefly discussed the various treatment options for patients with urolithiasis, and examined the timing of definitive treatment in the case of

urinary tract obstruction without infection. Our preliminary analysis suggests that the various definitive treatment options differ little in expected utility. However, an extensive sensitivity analysis of the relevant probabilities is beyond the scope of this article. Furthermore, a number of other issues should be included in an analysis of treatment for urolithiasis, such as pain, quality of life, financial costs of treatment and days lost due to morbidity. Thus, the analysis of treatment of ureteral stones could be the subject of a future paper.

Another clinical decision problem related to the one addressed is that of urinary tract obstruction caused by malignant disease. The employed drainage procedures are similar, except that an internal stent is preferred in such cases, which can be placed either via an antegrade or retrograde approach. Long-term complications, such as encrustation and obstruction of the stent, play an important role in this decision problem. However, the major issue is the quality of life for the short period of time that remains for the patient, and the preference of the patient to die of uraemia versus another direct cause of death.

In spite of the mentioned limitations our model suggests that the difference in benefits and risks between PCN and RUS are small. Both are low risk procedures and both are effective. Thus, the main point made in our analysis concerns not which of these two procedures should be performed, but rather whether intervention is indicated and when it should be performed. Whether intervention is indicated depends mainly on the probability of obstruction and pyonephrosis at presentation. Our model suggests that, faced with evidence of acute urinary tract obstruction, any sign of pyonephrosis is an indication for prompt drainage. If signs of obstruction without pyonephrosis exist, observing the patient for 5 to 15 days seems worthwhile to give the stone a chance to pass spontaneously. If the probability of obstruction is above 0.08 and no passage occurs within 5 to 15 days, our model suggests intervention at that time.

In a small fraction of cases obstruction of the urinary tract exists without dilatation of the collecting system, and the diagnosis of obstruction may be missed. Furthermore, many clinicians may wait until dilatation develops before performing PCN, because dilatation improves the likelihood of success. The literature, however, suggests that the probability of successfully placing a PCN tube in an undilated tract is only slightly lower than that in a dilated tract. Our model suggests that an undilated tract should not be a reason for postponing PCN if pyonephrosis is likely because the high risk of sepsis and renal impairment outweighs the small decrease in likelihood of a successfully placed PCN tube.

Urosepsis is frequently associated with an increased bleeding tendency, either due to thrombocytopenia or due to a derangement of coagulation factors. If a patient with signs of pyonephrosis has an increased bleeding tendency, RUS is preferred over PCN because haemorrhage is one of the major complications of PCN which may lead to death or loss of the kidney. However, if RUS fails, or if RUS is unavailable, PCN should be performed because, in spite of the increased bleeding tendency, the risk of PCN is smaller than the risks of medical or surgical management. Thus, an increased bleeding tendency constitutes a relative rather than an absolute, contraindication for PCN.

Although intuitively a solitary kidney may seem a contraindication to performing PCN, our model suggests that a solitary obstructed infected kidney should be drained mainly because if the affected kidney loses its function, the patient will develop renal failure and will require dialysis.

In conclusion, acute urinary tract obstruction with pyonephrosis is a remediable cause of urosepsis and renal failure. Percutaneous nephrostomy and retrograde ureteral stenting are both low risk drainage procedures, effective in urinary tract obstruction. If obstruction and pyonephrosis are likely, percutaneous nephrostomy should not be postponed because of a non-dilated tract: the probability of successful placement of a nephrostomy tube in a non-dilated tract is only slightly lower compared to a dilated tract. In patients with an increased bleeding tendency, retrograde ureteral stenting is preferred to percutaneous nephrostomy. However, if placement of the stent fails, or if stenting is unavailable, percutaneous nephrostomy may be attempted in spite of the increased risk. In a patient with pyonephrosis and a solitary functioning kidney, intervention is preferred over medical management because the risk of sepsis and the risk of loss of renal function with medical management exceeds the risk of intervention. In the presence of urinary tract obstruction without signs of pyonephrosis, postponing intervention between 5 to 15 days is appropriate, giving the stone a chance to pass spontaneously.

IX. REFERENCES

1. Andriole GL, Bettman MA, Garnick MB, Richie JP. Indwelling double-J ureteral stents for temporary and permanent urinary darinage: experience with 87 patients. J Urol 1984; 131: 239-241.

2. Androulakis PA. Pyonephrosis: a critical review of 131 cases. Br J Urol 1982; 54: 89-92.

3. Ansong K, Smith AD. Emergency Management of obstructive uropathy. Urol Clin North Am 1983; 10/1: 161-173.

4. Arafa NM, Fathi MM, Safwat M, Moro H, Torky H, Kenawi M, Abdel-Wahab M. Accuracy of ultrasound in the diagnosis of nonfunctioning kidneys. J Urol 1982; 128: 1165-1169.

5. Beck JR, Kassirer JP, Pauker SG. A convenient approximation of life expectancy (the "DEALE"). I. Validation of the method. Am J Med 1982; 73: 883-888.

6. Beck JR, Pauker SG, Gotlieb JE, Klein K, Kassirer JP. A convenient approximation of life expectancy (the "DEALE"). II. Use in Medical Decision-Making. Am J Med 1982; 73: 889-897.

7. Beck JR, Pauker SG. The Markov process in medical prognosis. Med Decis Making 1983; 3: 419-458.

8. Brenner BM, Rector FC. The Kidney, 3 ed, Saunders Company, Philadelphia. 1986; 1443-1483.

9. Coe FL, Keck J, Norton ER. The natural history of calcium urolithiasis. JAMA 1977; 238: 1519-1523.

10. De Petriconi R, Egghardt G, Frohneberg D, Hautmann R. Duoble-J ureteral stents: method without complication? Journal d'Urologie 1987; 93: 259-263.

11. Drach GW. Transurethral Ureteral Stone Manipulation. In: Harrison LH, Kandel LB (editors). Techniques in Urologic Stone Surgery. Mount Kisco, NY: Futura Publishing Co., Inc., 1986.

12. Drach GW, Dretler S, Fair W, Finlayson B, Gillenwater J, Griffith D, Lingeman J, Newman D. Report of the United States Cooperative Study of extracorporeal shockwave lithotripsy. J Urol 1986; 135: 1127-1133.

13. DuPont HL, Spink W. Infections due to Gram-negative organisms: an analysis of 860 patients with bacteriemia at the University of Minnesota Medical Center, 1958-1966. Medicine 1969; 45: 307-332.

14. Earley LE, Gottschalk CW. Strauss and Welt's Diseases of the Kidney. 3 ed, 1979. Little, Brown and Co. Boston. p 884-885.

15. Ellenbogen PH, Scheible FW, Talner LB, Leopold GR. Sensitivity of gray scale ultasound in detecting urinary tract obstruction. AJR 1978; 130: 731-733.

16. Flam TA, Malone MJ, Roth RA. Complications of Ureteroscopy. Urol Clin North Am 1988; 15/2: 167-174.

17. Fowler JE. The renal perfusion/excretion determination renogram: a new method of individual renal function determination in obstructie renal disease. J Urol 1978; 119: 449-452.

18. Fritzche P. Antegrade Pyelography: Therapeutic Applications. Radiol Clin North Am 1986 24 4: 573-86.

19. Gillenwater JY, Westervelt FB, Vaughan ED, Howards SS. Renal function after release of chronic unilateral hydronephrosis in man. Kidney Int 1975; 7: 179-186.

20. Gillenwater JY. Clinical aspects of urinary tract obstruction. Semin Nephrol 1982; 2: 46-54.

21. Gonzalez-Serva L, Weinerth JL, Glenn JF. Minimal mortality of renal surgery. Urology 1977; V IX no 3: 253-55.

22. Hahn PF, Yoder IC. Fever and flank pain: suspected infection of the upper urinary tract in adults. In: Diagnostic Imaging. An Algorithmic Approach. Eisenberg RL (ed) JB Lippincott Company. Philadelphia. 1988.

23. Harrison GS. The management of pyonephrosis. Ann R Coll Surg Engl 1983; 65: 126-7.

24. Ho PC, Talner LB, Parsons CL, Schmidt JD. Percutaneous nephrostomy: experience in 107 kidneys. Urology 1980; 16: 532 -35.

25. Holm-Nielsen A, Jorgensen T, Mogensen P, Fogh J. The prognostic value of probe renography in ureteric stone obstruction. Br J Urol 1981; 53: 504-507.

26. Jeffrey RB, Laing FC, Wing VW, Hoddick W. Sensitivity of sonography in pyonephrosis: a reevaluation. AJR 1985; 144: 71-73.

27. Kerr WS. Effect of complete ureteral obstruction for one week on kidney function. J Appl Physiol 1954; 6: 762-772.

28. Kerr WS. Effect of complete ureteral obstruction in dogs on kidney function. Am J Physiol 1956; 184: 521-526.

29. Kinn AC. Renal function in idiopathic hydronephrosis. Scand J Urol Nephrol 1983; 17: 169-174.

30. Klahr S. Pathophysiology of obstructive nephropathy. Nephrology Forum, Kidney Int 1983; V 23: 414-426.

31. Koff SA. The diagnosis of obstruction in experimental hydroureteronephrosis: Mechanisms for progressive urinary tract dilatation. Invest Urol 1981; V 19, no 2: 85-88.

32. Kreger BE, Craven DE, Carling PC, McCabe WR. Gram-negative bacteremia III. Reassessment of etiology, epidemiology and ecology in 612 patients. Am J Med 1980; 68: 332-343.

33. Kreger BE, Craven DE, McCabe WR. Gram-negative bacteremia IV. Re-evaluation of clinical features and treatment in 612 patients. Am J Med 1980; 68: 344-355.

34. Lang EK, Price ET. Redefinitions of indications for percutaneous nephrostomy. Radiology 1983; 147: 419-426.

35. Lang EK. Percutaneous nephrostolithotomy and lithotripsy: a multi-institutional survey of complications. Radiology 1987; 162: 25-30.

36. Lee ET. Statistical methods for Survival Data Analysis. Lifetime Learning Publications, Belmont, California 1980.

37. Lee WJ, Smith AD, Cubelli V, Badlani GH, Lewin B, Vernace F, Cantos E. Complications of percutaneous nephrolithotomy. AJR 1987; 148: 177-180.

38. Maillet PJ, Pelle-Francoz D, Laville M, Gay F, Pinet A. Nondilated obstructive acute renal failure: diagnostic procedures and therapeutic management. Radiology 1986; 160(3): 659-62.

39. Marberger M, Stackl W. Surgical treatment of renal calculi. In: Schneider HJ, editor, Urolithiasis: Therapy and prevention. 1986; Springer-Verlag, Berlin.

40. Mardis HK, Kroeger RM, Hepperlen TW, Mazer MJ, Kammandel H. Polyethylene Double-Pigtail Ureteral Stents. Urol Clin North Am 1982; 9/1: 95-101.

41. McSherry CK, Glenn F. The incidence and causes of death following surgery for nonmalignant biliary tract disease. Ann Surg 1980; 191: 271-275.

42. Naidich JB, Rackson ME, Mossey RT, Stein HL. Nondilated obstructive uropathy: percutaneous nephrostomy performed to reverse renal failure. Radiology 1986; 160(3): 653-7.

43. O'Flynn JD. The treatment of ureteric stones. Br J Urol 1980; 52: 436-438.

44. Paulson DF. The Urinary System. In: Textbook of Surgery, Sabiston DC (editor) 13 edition, Saunders 1986; 1638-69.

45. Personal communication: Grannum R Sant, Department of Urology, New England Medical Center.

46. Personal communication: Jerome P. Kassirer, Department of Medicine, New England Medical Center.

47. Pfister RC, Newhouse JH. Interventional percutaneous pyeloureteral techniques II. Percutaneous nephrostomy and other procedures. Radiol Clin North Am 1979; V 17: 351-63.

48. Pfister RC. Percutaneous Procedures. In: Uroradiology: an integrated approach. Friedland GW, Filly R, Goris ML, ea (editors). Churchill Livingstone, New York, 1983. 265-297.

49. Platt R, Polk BF, Murdock B, Rosner B. Mortality associated with nosocomial urinary-tract infection. N Engl J Med 1982; 307: 637-642.

50. Pridgen WR, Woodhead DM, Younger RK. Alterations in renal function produced by ureteral obstruction. Determination of critical obstruction time in relation to renal survival. JAMA 1961; V 178 no 6: 563-564.

51. Provoost AP, Molenaar JC. Renal function during and after a temporary complete unilateral ureter obstruction in rats. Invest Urol 1981; V 18 no 4: 242-246.

52. Reznek RH, Talner LB.Percutaneous Nephrostomy. Radiol Clin North Am 1984; V 22 no 2: 393-406.

53. Robinson MRG, Cross RJ, Shetty MB, Fittall B. Bacteriaemia and bacteriogenic shock in district hospital urological practice. Br J Urol 1980; 52: 10-14.

54. Saltzman B. Ureteral Stents. Indications, variations, and complications. Urol Clin North Am 1988; 15/3: 481-491.

55. Sandegard E. Prognosis of stone in the ureter. Acta Chirurgica Scandinavica 1956; Suppl 219.

56. Seidenfeld SM, Luby JP. Urologic Sepsis. Urol Clin North Am 1982; 9/2: 259.

57. Sinclair D, Wilson S, Toi A, Greenspan L. The evaluation of suspected renal colic: ultrasound scan versus excretory urography. Annals of Emergency Medicine 1989; 18:5 556-559.

58. Singer RB, Levinson L. Medical Risks: Patterns of mortality and survival. Lexington Books, DC Heath and Co, Lexington, Massachusetts, 1976. Tables 6-14, 6-21 and 6-27.

59. Sonnenberg FA, Wong JB, Pauker SG. Modelling time-dependent parameters with variable starting points in a Markov cohort simulation with limited memory. (Abstract) Med Decis Making 1988; 8: 350.

60. Sosa RE and Vaughan ED. Complications of Ureteroscopy. AUA Update Series, 1988 American Urological Association, Houston, Texas. VII: 35.

61. Stables DP, Ginsberg NJ, Johnson ML. Percutaneous Nephrostomy: A series and review of the literature. AJR 1978; 130: 75-82.

62. Stamm WE, Turck M. Urinary tract infection, pyelonephritis, and related conditions. In: Harrison's Principles of Internal Medicine. Petersdorf RG, Adams RD, Braunwald E, Isselbacher KJ, Martin JB, Wilson JD (editors). 10 ed. 1983; 1649-1655.

63. Subramanyam BR, Raghavendra BN, Bosniak MA, Lefleur RS, Rosen RJ, Horii SC. Sonography of pyonephrosis: a prospective study. AJR 1983; 140: 991-993.

64. Sullivan NM, Sutter VL, Mims MM, Marsh VH, Finegold SM. Clinical aspects of bacteriemia after manipulation of the genitourinary tract. J Infect Dis 1973; 127: 49-55.

65. Thornbury JR, Parker TW. Ureteral calculi. Semin Roentgenol 1982; 17/2: 133-139.

66. Vaughen ED, Gillenwater JY. Recovery following complete chronic unilateral ureteral occlusion: functional, radiographic and pathologic alterations. J Urol 1971; 106: 27-34.

67. Wegenke JD. Exchange/retrograde ureteral stent set. J Urol 1988; 140: 550-551.

68. Weil MH, Shubin H, Biddle M. Shock caused by Gram-negative microorganisms. Ann Int Med 1964; 60: 384-399.

69. Wyker AT, Ritter RC, Marion DN, Gillenwater. Mechanical factors and tissue stresses in chronic hydronephrosis. Invest Urology 1979; V 17 no 6: 430-436.

70. Yoder IC, Lindfors KK, Pfister RC. Diagnosis and treatment of pyonephrosis. Radiol Clin North Am 1984; 22(2): 407-14.

71. Yoder IC, Pfister RC, Lindfors KK, Newhouse JH. Pyonephrosis: imaging and intervention. AJR 1983; 141(4): 735-40.

Chapter V

THEORY of ROC ANALYSIS

I. Bayesian thinking and likelihood ratios
II. What is an ROC curve?
III. Sensitivity, specificity and ROC curves
IV. Obtaining data for an ROC analysis
V. ROC curve models and indices
1. Nonparametric estimates: the Mann Whitney method
2. Parametric estimates: ROC curves based on underlying distributions
VI. Comparing ROC curves
VII. The optimal operating point
VIII. Workup bias or verification bias
1. The basic assumption and an equivalent assumption
2. An example: ultrasound done for appendiceal disease.
3. Correcting ROC curves for verification bias
4. The relationship between the corrected and apparent likelihood ratio
5. How do the corrected P(T+|D+) and P(T+|D-) relate to the apparent P(T+|D+,V+) and P(T+|D-,V+)?
6. Implications of correcting for verification bias
IX. Uninterpretability bias
X. References

Receiver operating characteristic (ROC) methodology concerns the evaluation of diagnostic test performance and, thus, is important in diagnostic radiology. During the last 10 years an increasing interest has been shown by radiologists to use ROC analysis. In section I of this chapter the principle of Bayesian thinking is discussed in terms of likelihood ratios. A new term is defined, namely the "interval likelihood ratio", because the term "likelihood ratio" used in the literature refers to two different entities. In sections II to VII the basic concepts of ROC methodology are explained. The remaining part of the chapter deals with two selected methodological issues. Verification bias is discussed at length because it plays an important role in chapters VII and VIII. The correction method for verification bias was introduced by Begg et.al. (2,10), however, I introduce an equivalent but easier method (section VIII.2 and VIII.3). Moreover, the newly introduced method facilitates the correction for verification bias for a fully stratified data set in spite of small cell frequencies, by using a logistic regression analysis to model the probability of verification, which is utilized in the ROC analysis in chapter VIII. Furthermore, I derive an equation to show how the corrected and apparent rates relate, illustrating how verification bias affects the calculated rates depending on the ratio of verification of positive and negative test results.

I. Bayesian thinking and likelihood ratios

In the initial phase of the diagnostic process a clinician integrates information from the history and physical examination to list a differential diagnosis, listing the most probable diagnosis up front. Further workup is a sequential process of adjusting the existing clinical impression with the new data from the performed diagnostic tests. Each piece of new information will be considered in the light of two questions:

1) How often have I seen this finding in patients who had the disease?
2) How often have I seen this finding in patients who did not have the disease?

If the answers to these questions are more or less equal, then the new finding is useless. If the finding is seen often in patients who have the disease and seldom in patients without the disease, then the new finding is an argument in favor of the diagnosis. Vice versa, if the finding is seldom seen in patients who have the disease and often in patients without the disease, the new finding is an argument against the diagnosis. The existing clinical impression is thus adjusted using the information from the new finding.

The above train of thought can be summarized in probability notation in the form of Bayes rule. The initial, or existing, impression is expressed as the probability of disease $P(D^+)$, or as the odds favoring disease, with

$$odds = \frac{P(D^+)}{1 - P(D^+)} \qquad (Eq.1)$$

The probability of the new finding T in patients with the disease is, in probability notation, $P(T|D^+)$, also known as the "true positive rate", and in those without the disease $P(T|D^-)$, the "false positive rate". (Note that the "true and false positive rates" are actually observed relative frequencies, used to estimate probabilities.) The likelihood ratio LR is:

$$LR = \frac{P(T \mid D^+)}{P(T \mid D^-)} \qquad (Eq.2)$$

A likelihood ratio LR = 1 implies that the new finding is seen equally often in patients with and without the disease. If the LR > 1 the new finding argues in favor of the diagnosis. If the LR is infinity the new finding is pathognomic for the disease. If the LR < 1 the new finding argues against the diagnosis and if the LR = 0 the new finding excludes the disease. Adjusting the existing clinical impression with the information from the new finding is equivalent to multiplying the odds with the LR:

$$adjusted\ odds = odds \times LR \qquad (Eq.3)$$

Thus far we have discussed a dichotomous finding, ie. the finding is either present or absent. Analogously, a test result may be dichotomous, ie. the test result is either positive or negative. However, often a test result is on a scale, which can be either ordinal or continuous. A radiologist often intuitively expresses his/her confidence in the diagnosis on an ordinal scale. Results of biochemical tests are usually given on a continuous scale, which may be reduced to an ordinal scale by categorizing the test variable. If the test variable is ordinal or continuous, we have to consider many test results R_i, where i can be any value from 1 to the number of categories. A test result on an ordinal scale is simply a generalization of a dichotomous test result. A test result on a continuous scale can be considered a result on an ordinal scale with an infinite number of very narrow categories.

The term "likelihood ratio" has been used in the literature to mean two different quantities. We propose making a distinction between the "interval likelihood ratio" (interval LR) and the "likelihood ratio" (LR). In probability notation, the interval likelihood ratio (interval LR_i) of the test result R_i is:

$$intervalLR_i = \frac{P(R_i \mid D^+)}{P(R_i \mid D^-)} \qquad (Eq.4)$$

where $P(R_i|D^+)$ is the probability of the test result R_i given disease, and $P(R_i|D^-)$ is the probability of the test result R_i given no disease.

The likelihood ratio LR_i is the ratio of the true positive rate and false positive rate at the cutoff value i. LR_i expressed in probability notation is:

$$LR_i = \frac{\sum_{k=1}^{i} P(R_k \mid D^+)}{\sum_{k=1}^{i} P(R_k \mid D^-)} \qquad (Eq.5)$$

which is equivalent to equation 2.

The distinction between the likelihood ratio and interval likelihood ratio is important in ROC methodology, as will become apparent in this chapter.

II. What is an ROC curve?

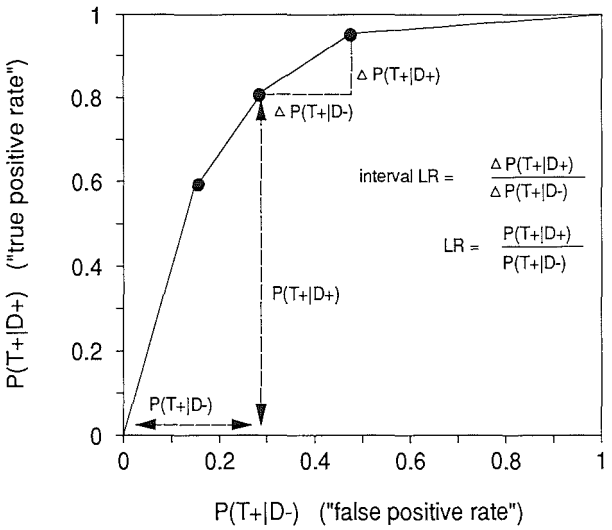


Figure 1. An ROC curve, the likelihood ratio (LR) and interval LR for an ordinal test result.

An receiver operating characteristic (ROC) curve is a plot of the true and false positive rates, that is, $P(T^+|D^+)$ is plotted as function of $P(T^+|D^-)$, for different cutoff values of the test result (8,11,13,16,22). Each pair of $\{P(T^+|D^+),P(T^+|D^-)\}$ is an operating point of the ROC curve. In the ROC plot, the likelihood ratio (LR) is the ratio of the values of the coordinates. The interval likelihood ratio (interval LR) is the marginal true positive rate divided by the marginal false positive rate between two adjacent points on the ROC curve (figure 1). The closer the operating points on the ROC curve are, the smoother the curve and the closer the interval LR becomes to the slope of the curve. For a continuous test variable the interval likelihood ratio is equal to the slope of the ROC curve at any particular operating point (figure 2). Going from (0,0) to (1,1) the interval LR and thus the slope of the ROC curve, usually decreases monotonically, that is, for increasing $P(T^+|D^+)$, the marginal increase of $P(T^+|D^+)$, compared to the marginal increase of $P(T^+|D^-)$, diminishes.

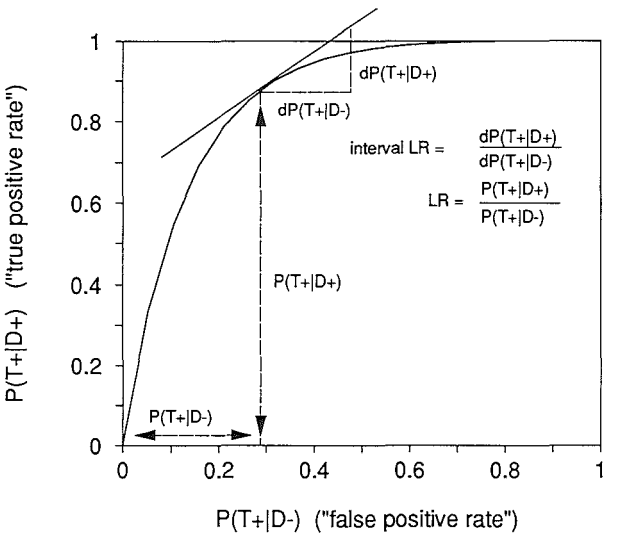


Figure 2. An ROC curve, the likelihood ratio (LR) and interval LR for a continuous test result.

III. Sensitivity, specificity and ROC curves

In describing the performance of diagnostic test systems various indices are in use (16). Sensitivity and specificity are indices well known in radiological literature. Using a 2 x 2 table (table 1) the terms are easily understood. Sensitivity is the probability that the test result is positive conditional on the fact that the patient has the disease (also known as the true positive rate) (table 1b). Specificity is the probability that the test is negative conditional on the fact that the patient does not have the disease (also known as the true negative rate) (table 1b). The false positive rate equals one minus the specificity (table 1b).

Sensitivity and specificity, and thus the true and false positive rates, depend not only on the capacity of the test to distinguish diseased from non-diseased, but also on the decision criterion chosen by the observer. In other words, the imaging process itself determines part of overall test performance whereas the observer of the image (apart from his/her ability to do the task at hand) has to decide whether to "label" the patient as diseased or not (16). The decision to label a patient as diseased depends not only on seeing signs of the disease on the image, but also on the consequences of making the particular diagnosis. On average the expected benefits of making the diagnosis, rightly or wrongly, have to outweigh the expected costs. In general, if the disease

Table 1a. 2 x 2 table of true and false positive and true and false negative frequencies in the evaluation of a diagnostic test.

	TEST RESULT		
FINAL DIAGNOSIS	test positive	test negative	total
disease	$n.P(T^+,D^+)$	$n.P(T^-,D^+)$	$n.P(D^+)$
no disease	$n.P(T^+,D^-)$	$n.P(T^-,D^-)$	$n.P(D^-)$
total	$n.P(T^+)$	$n.P(T^-)$	n

Table 1b. Sensitivity and specificity, true and false positive rates and predictive values.

true positive rate = sensitivity	=	$P(T^+\|D^+)$	=	$n.P(T^+,D^+)/n.P(D^+)$
false positive rate = 1-specificity	=	$P(T^+\|D^-)$	=	$n.P(T^+,D^-)/n.P(D^-)$
specificity	=	$P(T^-\|D^-)$	=	$n.P(T^-,D^-)/n.P(D^-)$
predictive value positive given a positive test result	=	$P(D^+\|T^+)$	=	$n.P(T^+,D^+)/n.P(T^+)$
predictive value negative given a negative test result	=	$P(D^-\|T^-)$	=	$n.P(T^-,D^-)/n.P(T^-)$

is common, if it is associated with a bad outcome when the diagnosis is missed, and if it has an affordable low risk treatment, the observer will use lenient criteria to make the diagnosis. The result will be many true positive diagnoses (high sensitivity) but also many false positives (low specificity). Vice versa, if the disease is sporadic, has a good prognosis if untreated, and the available treatment has a high risk, the observer will use strict criteria resulting in a low true positive rate and low false positive rate. Thus, with an increase in true positive rate, the false positive rate also increases. This phenomenon is analogous to shifting a cutoff value of a test variable, of for example a biochemical test, for which the normal and abnormal populations have overlapping distributions (figure 3). Although in radiology we usually do not have a numeric variable, we can conceive of the observer's subjective judgement of the likelihood of disease as having a continuous scale of measurement, and so the radiologist can express his/her confidence in the diagnosis on a rating scale.

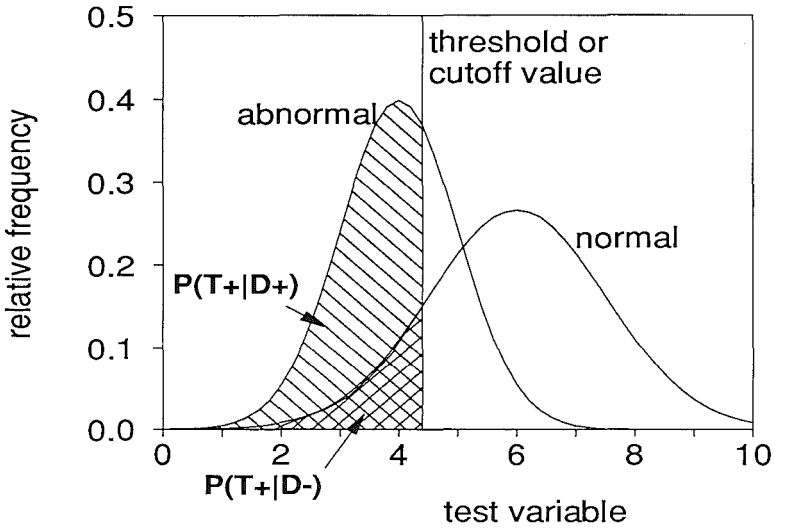


Figure 3. Overlapping distributions of normal and abnormal test results. When the threshold, or cutoff value, is shifted to higher values, both true and false positive rates increase.

Receiver operating characteristic (ROC) methodology provides an alternative method to sensitivity and specificity in describing a test system's performance, independent of the ultimately chosen confidence threshold (16). A ROC curve is a plot of true positive rates as a function of the corresponding false positive rates (figure 4). The ROC curve plots pairs of true positive and false positive rates at different confidence thresholds, thereby evaluating overall system performance, unaffected by the decision criterion.

The points (0,0) and (1,1) are inherent to all ROC curves, simply representing the two situations that all patients are labelled non-diseased and diseased, respectively. Points on the curve in the lower left corner (figure 4) represent situations in which strict criteria are used to make the diagnosis: there will be few false positives, but also few true positives. Points on the curve in the upper right hand corner represent situations in which lenient criteria are used to make the diagnosis: many true positive diagnoses are made but also many false positives.

Curves that go up steeply from (0,0) and reach near (0,1) describe test systems with good discriminating power. Curves close to the diagonal have little or no discriminating power (figure 4). The area under the ROC curve gives some indication of how good the test performance is, independent of the chosen operating point on the curve. An area of 1 represents the ideal, while an area of 0.5 indicates a test with no discriminating power.

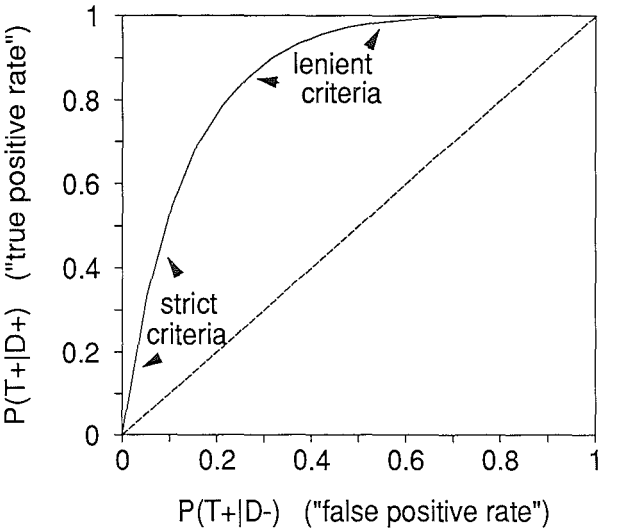


Figure 4. An ROC curve (solid line). The dashed diagonal represents the ROC curve of a test with no discriminating power.

Receiver operating characteristic analysis (in the past also known as relative operating characteristic analysis) was developed in signal detection theory (7,8,11,22). The technique is also used in psychology, polygraph lie detection and weather forecasting. In the last 10 years the technique has increasingly been applied to medical imaging (16).

IV. Obtaining data for an ROC analysis

If the diagnostic test result is a value, such as with biochemical tests, these values are used to determine if a test is positive or negative for any particular cutoff point. With diagnostic imaging this approach is usually not possible, because no numerical value indicating the presence or absence of disease is produced. However, the reader of the image can (and in the clinical routine usually does) express his/her confidence in the diagnosis.

Table 2. The true and false positives for two of the four thresholds: a) for the threshold between equivocal and probably negative and b) for the threshold between equivocal and probably positive, for an hypothetical example.

a)

FINAL DIAGNOSIS	positive	probably positive	equi-vocal	probably negative	negative
positive	6	24	4	6	2
	true positives			false negatives	
negative	8	53	79	115	361
	false positives			true negatives	

b)

FINAL DIAGNOSIS	positive	probably positive	equi-vocal	probably negative	negative
positive	6	24	4	6	2
	true positives		false negatives		
negative	8	53	79	115	361
	false positives		true negatives		

One method of obtaining points on the ROC curve of an imaging test is with the "yes-no method": this would imply presenting the same set of images to the observer several times, in which the observer would be requested to choose a different confidence threshold each time the set of images is read (16). Clearly, this is not a very efficient method and many radiologists will soon lose interest in the research project.

A more convenient method is the "rating method" (16). Every time an observer reads an image, he marks his confidence in the diagnosis on a rating scale. For example, a scale often used is:

1/ definitively positive
2/ probably positive
3/ equivocal
4/ probably negative
5/ definitively negative

Dividing the scale into five to seven categories is usually enough and more categories will not give more information (19), provided that the results are more or less evenly spread over the categories. (The additional categories on a seven category scale could, for example, be "very probably positive" and "very probably negative".)

The true and false positive rates are subsequently calculated by shifting the threshold for the diagnosis (table 2), which is analogous to shifting the cutoff value of a test variable as shown in figure 3. Each possible threshold defines a combination of false and true positive rates, or equivalently, of sensitivity and specificity.

V. ROC curve models and indices

The area under the ROC curve, and the standard error of the area, are the customarily used measures to describe overall test performance. The area under the ROC curve is equivalent to the probability that a randomly chosen pair of normal and abnormal images will be correctly diagnosed, in other words, that they will be correctly ranked on the confidence scale (13). Various models of ROC curves and the related indices exist (8). These can be divided into two types: the nonparametric methods and the parametric methods.

1. Nonparametric estimates: the Mann-Whitney method

The most convenient way of calculating the area under the ROC curve is by connecting the calculated points and adding the areas of the trapezoids underneath each part of the curve. This is a nonparametric method, that is, no assumptions are made as to the underlying distributions of test results. The method is equivalent to the Mann-Whitney-Wilcoxon U statistic (1). The U statistic is normally used to test whether the value of a quantitative variable is generally larger in one population compared to another population. Its value in this context is that in addition to providing a formula for calculating the area, we can also easily determine its standard error according to the Hanley-McNeil algorithm (13). The Mann-Whitney method is adequate provided enough data points are present and the points are spread along the curve (5). The area under the ROC curve is slightly underestimated with this method, but if comparing tests is the major issue, the underestimation of the area is of little concern.

2. Parametric estimates: ROC curves based on underlying distributions

Parametric methods are based on the assumption that the test results conform to some well defined underlying distribution. The most commonly used parametric method is the one introduced by Dorfman and Alf (7), which assumes that the underlying distributions of the test results are Gaussian, or normal. Although the distributions of test results are often not Gaussian,

it has recently been shown that the binormal ROC model gives accurate results for many non-Gaussian distributions (14). Other underlying distributions include chi-squared, lognormal and the negative exponential (8,9).

When using the binormal ROC model the basic procedure is to convert the true and false positive rates to their corresponding normal deviate values. The ROC curve plotted on binormal deviate axes, is a straight line (21,22). A maximum likelihood estimation algorithm is conventionally used to calculate the slope and intercept of the line (7). However, in many circumstances a simpler, non-iterative least squares regression will produce an adequate fit. This is especially useful since in a number of situations the maximum likelihood estimation procedure will fail to converge (15).

VI. Comparing ROC curves

In comparing the area under two ROC curves a two-sided paired t-test is used. If the test results were derived from the same set of patients, one should take into account the correlation between test results (12). The z statistic is calculated using the formula:

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$$

where A_1 and A_2 are the areas under the ROC curves of the two tests, SE_1 and SE_2 are the standard errors of the areas, and r is the correlation coefficient as found in table I of reference 12, a coefficient representing the correlation between the areas under two ROC curves derived from the same cases. r is based on the average area of the two ROC curves and the average of r_N and r_D, the Kendall tau (for ordinal scales) or Pearson product moment correlation (for continuous scales) between the two test results, calculated separately for non-diseased (r_N) and diseased (r_D) patients.

VII. The optimal operating point

The optimal operating point is the point on the ROC curve which we choose as working point, taking into account the utility of true and false positive, and true and false negative outcomes. The utility structure should consist of costs, risks and benefits.

The optimal operating point is calculated as follows (20). If the utilities of the true and false positive outcomes are $U(T^+,D^+)$ and $U(T^+,D^-)$, and the utilities of the true and false negative outcomes are $U(T^-,D^-)$ and $U(T^-,D^+)$, then the overall utility U of performing the test is (figure 5):

$$U = P(D^+)\cdot[P(T^+\mid D^+)\cdot U(T^+,D^+)+(1-P(T^+\mid D^+))\cdot U(T^-,D^+)]$$
$$+(1-P(D^+))\cdot[P(T^+\mid D^-)\cdot U(T^+,D^-)+(1-P(T^+\mid D^-))\cdot U(T^-,D^-)]$$

where
$P(D^+)$ is the prior probability of the disease
$P(T^+|D^+)$ is the true positive rate and
$P(T^+|D^-)$ is the false positive rate

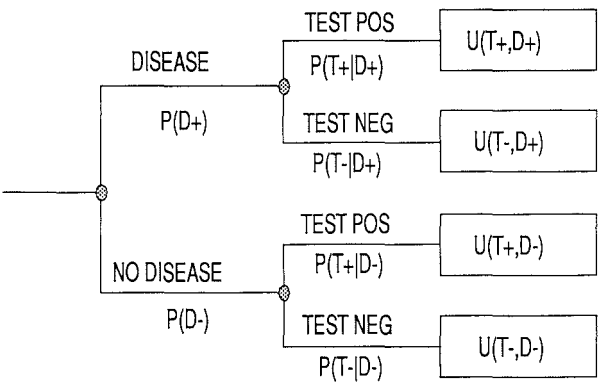


Figure 5. Calculating the optimal operating point: the tree structure of performing a test.

The optimal operating point is the point $(P(T^+|D^-),P(T^+|D^+))$ such that the utility of the tree is maximized. In other words, we wish to maximize U with respect to both $P(T^+|D^+)$ and $P(T^+|D^-)$. This implies setting the first derivative of U, with respect to $P(T^+|D^+)$ and $P(T^+|D^-)$, to zero:

$$\frac{\delta U}{\delta P(T^+\mid D^+)}=0 \qquad \text{and} \qquad \frac{\delta U}{\delta P(T^+\mid D^-)}=0$$

Taking the first derivative of U with respect to $P(T^+|D^+)$ we find:

$$\frac{\delta U}{\delta P(T^+\mid D^+)}$$
$$=P(D^+)\cdot(U(T^+,D^+)-U(T^-,D^+))+(1-P(D^+))\cdot\left(U(T^+,D^-)\cdot\frac{\delta P(T^+\mid D^-)}{\delta P(T^+\mid D^+)}-U(T^-,D^-)\cdot\frac{\delta P(T^+\mid D^-)}{\delta P(T^+\mid D^+)}\right)$$
$$=P(D^+)\cdot(U(T^+,D^+)-U(T^-,D^+))+(1-P(D^+))\cdot(U(T^+,D^-)-U(T^-,D^-))\cdot\frac{\delta P(T^+\mid D^-)}{\delta P(T^+\mid D^+)}$$

Setting

$$\frac{\delta U}{\delta P(T^+ \mid D^+)} = 0$$

we derive

$$\boxed{\frac{\delta P(T^+ \mid D^+)}{\delta P(T^+ \mid D^-)} = \frac{1 - P(D^+)}{P(D^+)} \cdot \frac{U(T^-, D^-) - U(T^+, D^-)}{U(T^+, D^+) - U(T^-, D^+)}}$$

Note that taking the first derivative of U with respect to $P(T^+|D^-)$ gives exactly the same result. As explained in sections I. and II., $\delta P(T^+ \mid D^+)/\delta P(T^+ \mid D^-)$ is the slope of the ROC curve, which equals the interval likelihood ratio. Using the above equation, we calculate the slope, and thus the interval likelihood ratio, at the optimal operating point, from which follows the value, or category, of the test result that should be used to maximize the utility of the test.

VIII. Workup bias or verification bias

Workup bias or verification bias occurs when only a subset of all tested subjects are selected for further workup and verification of the test result (4,18). Often the test result, possibly in combination with other test results and clinical features, influence the decision to do a workup or not and thus the test result will in part determine the chance that the diagnosis will be verified or not. Biased selection for verification can have an effect on the calculated test characteristics if we simply ignore the unverified cases. If we assume that the decision to verify the diagnosis depends only on the test results and clinical information, we can calculate unbiased estimates of the test parameters (2,10). First, the basic assumption is discussed in more detail. In paragraph 2 an example is presented.

1. The basic assumption and an equivalent assumption

The basic assumption for the correction of verification bias, as introduced by Begg and Greenes (2), is:

$$P(V^+ \mid R_i) = P(V^+ \mid R_i, D)$$

In words: the probability of verification is conditionally independent of true disease status D. The decision to verify depends only on the test result R_i. The influence of true disease status on selection for verification is only through the effect the disease has on the test result.

In practice, selection for verification will be biased not only by the test result but also by clinical features, such as signs, symptoms, age or sex. (The term "test" should be considered in a broad sense. Clinical evaluation including, signs and symptoms, could also be called a "test".) To correct for verification bias produced by multiple factors, eg. test result R_i and clinical information X_j, the basic assumption for the correction is:

$$P(V^+ \mid R_i, X_j) = P(V^+ \mid R_i, X_j, D)$$

In words: the probability of verification does not depend on the true disease status D, but only on the test result R_i and clinical information X_j.

Although not immediately obvious, the above assumption is equivalent to saying that the predictive values are unaffected by verification bias (2). In probability notation this can be proven as follows. Assuming

$$P(V^+ \mid R_i, X_j) = P(V^+ \mid R_i, X_j, D)$$

then

$$\frac{P(V^+, R_i, X_j)}{P(R_i, X_j)} = \frac{P(V^+, R_i, X_j, D)}{P(R_i, X_j, D)}$$

$$\frac{P(D, R_i, X_j)}{P(R_i, X_j)} = \frac{P(D, V^+, R_i, X_j)}{P(V^+, R_i, X_j)}$$

and therefore

$$P(D \mid R_i, X_j) = P(D \mid R_i, X_j, V^+)$$

which means that the probability of disease, given test result R_i and clinical information X_j, does not depend on whether the diagnosis is verified or not.

2. An example: ultrasound done for appendiceal disease.

Consider the following example: Puylaert et.al. reported the results of a study on ultrasound (US) examination in 111 patients in the diagnosis of possible acute appendicitis, perforating appendicitis or appendiceal mass (17). Although the authors did not correct the data for verification bias, they did report the frequency of unverified ultrasound results. In 83 cases a definite diagnosis was made at operation or laparoscopy or during further radiological and clinical workup. In 28 cases no definite diagnosis could be made. The results are given in table 3. The authors of the article have distinguished between unequivocal and dubious results, and a non-visualized appendix. Their reported sensitivity of 0.75 and specificity of 1 is based on including among the positive test results only those with unequivocal signs on US, and including among

the negative results those with a dubious US or a non-visualized appendix. For the sake of the argument we will consider both unequivocal and dubious test results as positive test results, and a non-visualized appendix as a negative test result.

Table 3. Results of a study on ultrasound for appendiceal disease, a) original data and b) data as used for illustrative purposes.

a)

	US unequivocal	US dubious	appendix not visualized	total
appendiceal disease	39	3	10	52
no appendiceal disease	0	3	28	31
unverified	4	1	23	28
total	43	7	61	111

b)

FINAL DIAGNOSIS	US positive	US negative	total
appendiceal disease	42	10	52
no appendiceal disease	3	28	31
total verified	45	38	83
unverified	5	23	28
total	50	61	111

Considering the verified cases only, we calculate the true positive rate (sensitivity), the false positive rate (1-specificity), the likelihood ratio and the predictive values as follows:

true positive rate (sensitivity) = $P(T^+|D^+)$ = 42/52 = 0.81
false positive rate (1-specificity) = $P(T^+|D^-)$ = 3/31 = 0.10
likelihood ratio = LR = .81/.10 = 8.1

predictive value positive
given a positive test = $P(D^+|T^+)$ = 42/45 = 0.93
predictive value negative
given a negative test = $P(D^-|T^-)$ = 28/38 = 0.74

From table 3b it is clear that far more negative test results are left unverified than positive test results, that is 23 versus 5. Intuitively this makes sense since if the result of the US is negative the surgeons will not be keen to operate and further workup for appendiceal disease will not be performed. The probability that the diagnosis is verified can be expressed conditional on the test result as follows:

probability of verification
given a positive test = $P(V^+|T^+)$= 45/50 = 0.90
given a negative test = $P(V^+|T^-)$= 38/61 = 0.62

Now the question is if we know that 90% of the positive test results and 62% of the negative test results are verified, can we estimate what the results would have been had all patients been verified. We can, provided we assume that verification depends only on the test result. If a sample is a fraction f of a source population, the source population will be 1/f times the size of the sample. Thus, to estimate the cell frequencies had all patients been verified, we divide the observed cell frequencies among the verified patients by the probability of verification given the test result. That is, we divide the positive test results by 0.90 (= $P(V^+|T^+)$) and the negative test results by 0.62 (= $P(V^+|T^-)$). The estimated results, had all patients been verified, are given in table 4. Written in probability notation, for $D = D^+$ and $D = D^-$, the calculations are:

$$n \cdot P(D, T^+) = n \cdot P(D, T^+, V^+) \cdot \frac{1}{P(V^+ \mid T^+)}$$

$$n \cdot P(D, T^-) = n \cdot P(D, T^-, V^+) \cdot \frac{1}{P(V^+ \mid T^-)}$$

The correction method is equivalent to distributing the unverified cases over the D^+ and D^- classes in such a way that the predictive values stay the same. Important to remember when using this estimation procedure, is the underlying assumption that the probability of verification depends only on the test result, that is, only indirectly on true disease status through the effect the disease has on the test result.

From the corrected table we calculate the true positive rate (sensitivity), the false positive rate (1-specificity), the likelihood ratio and the predictive values as follows:

true positive rate (sensitivity) = $P(T^+|D^+)$ = 46.7/63 = 0.74
false positive rate (1-specificity) = $P(T^+|D^-)$ = 3.3/48 = 0.07
likelihood ratio = LR = .74/.07 = 10.6

Table 4. Cell frequencies corrected for verification bias, assuming that verification depends only on the test result.

	US positive	US negative	total
appendiceal disease	42/0.90 = 46.7	10/0.62 = 16.1	63
no appendiceal disease	3/0.90 = 3.3	28/0.62 = 45.0	48
total	50	61	111

predictive value positive
given a positive test = $P(D^+|T^+)$ = 46.7/50 = 0.93
predictive value negative
given a negative test = $P(D^-|T^-)$ = 45.0/61 = 0.74

Comparing the corrected test characteristics to those calculated from the verified patients we note that the true positive rate ($P(T^+|D^+)$) and false positive rate ($P(T^+|D^-)$ =1-specificity) are both lower than the group of verified patients suggested. In other words, the sensitivity is actually lower and the specificity is higher than the group of verified patients suggested[1]. However, the predictive values are the same for the verified patients and estimated population. The LR increases in this example, however, the LR may increase or decrease depending on the probability of disease among the verified cases and the actual probability of disease.

3. Correcting ROC curves for verification bias

The easiest method to correct ROC curves for verification bias is to estimate the frequency table had all patients been verified, as done in the example in paragraph 2. This method also holds for a 5 x 2 table, or for any number of dimensions and any number of categories. This is proven as follows.

[1]If only unequivocal test results are considered positive test results (as the authors of the article have done) the corrected sensitivity is 68% and corrected specificity 100%. Because the specificity is the maximum attainable, the corrected specificity will not be higher. (We chose to consider dubious test results as positive so that the effect on specificity would also become clear.)

Assume n is the total number of cases and D is the true disease status, either D^+ or D^-. From the observations we know the following:

$n \cdot P(D, R_i, X_j, V^+) =$ the number of verified cases with test result R_i and clinical information X_j and

$P(V^+ \mid R_i, X_j) =$ the probability of verification given test result R_i and clinical information X_j.

To correct the ROC curve we need to estimate

$n \cdot P(D, R_i, X_j) =$ the number with disease status D, test result R_i and clinical information X_j, had all patients been verified.

According to conditional probability theory:

$$n \cdot P(D, R_i, X_j) = n \cdot P(D \mid R_i, X_j) \cdot P(R_i, X_j)$$

From the basic assumption followed the equivalent assumption (paragraph 1)

$$P(D \mid R_i, X_j) = P(D \mid R_i, X_j, V^+)$$

and thus

$$\begin{aligned} n \cdot P(D, R_i, X_j) &= n \cdot P(D \mid R_i, X_j, V^+) \cdot P(R_i, X_j) \\ &= n \cdot \frac{P(D, R_i, X_j, V^+)}{P(R_i, X_j, V^+)} \cdot P(R_i, X_j) \\ &= n \cdot P(D, R_i, X_j, V^+) \cdot \frac{1}{P(V^+ \mid R_i, X_j)} \end{aligned}$$

and therefore:

$$\boxed{n \cdot P(D, R_i, X_j) = n \cdot P(D, R_i, X_j, V^+) \cdot \frac{1}{P(V^+ \mid R_i, X_j)}}$$

with

$n \cdot P(D, R_i, X_j) =$ the estimated cell frequency, had all patients been verified, of the cell R_i and X_j,

$n \cdot P(D, R_i, X_j, V^+) =$ the observed cell frequency, among verified patients, of the cell R_i and X_j and

$P(V^+ \mid R_i, X_j) =$ the probability of verification given R_i and X_j.

The derived formula holds for all i and j and for D^+ and D^-, so that all cell frequencies of the frequency table for the projected results, had all patients been verified, can be calculated.

The equation for the correction method described in the original papers on verification bias (2,10) is derived by substituting the assumption (see paragraph 1)

$$P(D \mid R_i, X_j) = P(D \mid R_i, X_j, V^+)$$

in Bayes theorem. The derived equation is:

$$P(R_i \mid D) = \frac{\sum_j P(R_i, X_j) \cdot P(D \mid R_i, X_j, V^+)}{\sum_i \sum_j P(R_i, X_j) \cdot P(D \mid R_i, X_j, V^+)}$$

It is more convenient correcting the frequency table of an ROC analysis by dividing by the probability of verification, as introduced in this thesis, than using the formula described in the original papers. The calculations of the ROC curve from the corrected frequency table are equivalent to the calculations for the verified group of patients, which can be done with a simple spreadsheet template (6) or any simple mathematics program. Furthermore, we obtain an estimate of the projected population from the corrected table, which helps understanding verification bias and facilitates presentation of the results. A simple example of correction for verification using this formulation is presented in chapter VII. Another advantage of the here presented formulation is its convenience in correcting for verification for a data set that is fully stratified, for example for many tests and/or many test results. A problem which often arises when correcting a fully stratified data set, is that the cell frequencies become very small, with the result that the calculated probabilities of verification for any particular cell are inaccurate. A solution to this problem is to perform a logistic regression analysis on the probability of verification, which takes into account all the available data. Subsequently, the frequency table can be corrected with the above presented method. An example of a logistic regression analysis of the probability of verification, and subsequent correction for verification, is given in chapter VIII.

4. The relationship between the corrected and apparent likelihood ratio

The relationship between the corrected and apparent LR is easily derived from the equivalent assumption stated in paragraph 1. The equivalent assumption stated that the predictive value is the same in the verified patients and after correction for verification bias. Thus, for test result R_i

$$P(D^+ \mid R_i) = P(D^+ \mid R_i, V^+)$$

From this follows (10)

$$LR \cdot \frac{P(D^+)}{1-P(D^+)} = LR^v \cdot \frac{P(D^+ \mid V^+)}{1-P(D^+ \mid V^+)}$$

with
LR = corrected likelihood ratio, had all patients been verified
LR^v = apparent likelihood ratio, among the verified patients
$P(D^+)$ = true prior probability of disease, either estimated or derived from another source
$P(D^+ \mid V^+)$ = apparent prior probability of disease

5. How do the corrected $P(T^+|D^+)$ and $P(T^+|D^-)$ relate to the apparent $P(T^+|D^+,V^+)$ and $P(T^+|D^-,V^+)$?

Although the effect of verification bias on the true and false positive rates has been discussed in the literature (2,10), the direction in which the bias affects the rates is not immediately obvious. A simple equation expressing the corrected rates in terms of the apparent rates has not been described before. In this section I derive the relevant equations and discuss their implications.

Consider a population of n patients consisting of $n \cdot P(D^+)$ diseased and $n \cdot P(D^-)$ non-diseased patients (figure 6), with $P(D^+)$ and $P(D^-)$ the probability of disease and non-disease respectively. A diagnostic test T, performed to detect disease D, has a dichotomous result with a true positive rate (or sensitivity) of $P(T^+ \mid D^+)$ and a false positive rate (or 1-specificity) of $P(T^+ \mid D^-)$. The probability of verification given a positive test result is $P(V^+ \mid T^+)$ and given a negative test result is $P(V^+ \mid T^-)$.

Among the group of verified patients there are $n \cdot P(T^+, D^+, V^+)$ true positive cases and $n \cdot P(T^-, D^+, V^+)$ false negative cases. The apparent true positive rate, calculated from the verified cases only, is therefore

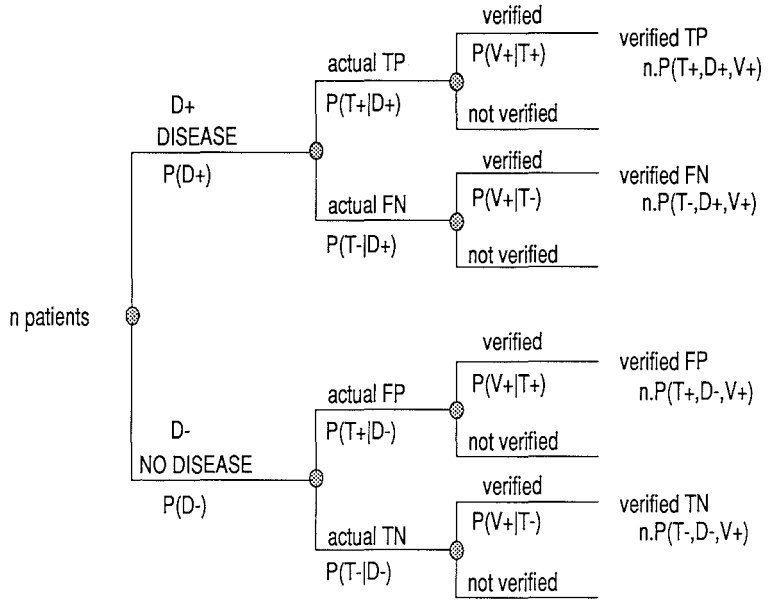


Figure 6. Illustration of verification bias.

$$P(T^+ \mid D^+, V^+) = \frac{n \cdot P(T^+, D^+, V^+)}{n \cdot P(T^+, D^+, V^+) + n \cdot P(T^-, D^+, V^+)}$$

Substituting the path probabilities, as defined above and shown in figure 6 we find

$$P(T^+ \mid D^+, V^+) = \frac{P(D^+) \cdot P(T^+ \mid D^+) \cdot P(V^+ \mid T^+)}{P(D^+) \cdot P(T^+ \mid D^+) \cdot P(V^+ \mid T^+) + P(D^+) \cdot P(T^- \mid D^+) \cdot P(V^+ \mid T^-)}$$

$$= \frac{P(T^+ \mid D^+) \cdot P(V^+ \mid T^+)}{P(T^+ \mid D^+) \cdot P(V^+ \mid T^+) + (1 - P(T^+ \mid D^+)) \cdot P(V^+ \mid T^-)}$$

$$= \frac{1}{1 + \left(\frac{1}{P(T^+ \mid D^+)} - 1 \right) \cdot \frac{P(V^+ \mid T^-)}{P(V^+ \mid T^+)}}$$

Let VR be the verification ratio with

$$VR = \frac{P(V^+ \mid T^+)}{P(V^+ \mid T^-)}$$

then

$$1 + \left(\frac{1}{P(T^+ \mid D^+)} - 1\right) \cdot \frac{1}{VR} = \frac{1}{P(T^+ \mid D^+, V^+)}$$

$$\frac{1}{P(T^+ \mid D^+)} - 1 = VR \cdot \left(\frac{1}{P(T^+ \mid D^+, V^+)} - 1\right)$$

and therefore

$$\boxed{P(T^+ \mid D^+) = \frac{1}{VR \cdot \left(\frac{1}{P(T^+ \mid D^+, V^+)} - 1\right) + 1}}$$

with

$P(T^+|D^+,V^+)$ = apparent true positive rate (among the verified patients),
$P(T^+|D^+)$ = corrected true positive rate (the unbiased estimate) and
VR = the verification ratio as defined above.

In a similar fashion we derive the equation for the false positive rate:

$$\boxed{P(T^+ \mid D^-) = \frac{1}{VR \cdot \left(\frac{1}{P(T^+ \mid D^-, V^+)} - 1\right) + 1}}$$

The derived equations are useful if we wish to calculate unbiased estimates of the $P(T^+|D^+)$ and $P(T^+|D^-)$ at one particular threshold. They give insight as to the relationship of the corrected $P(T^+|D^+)$ compared to the apparent $P(T^+|D^+,V^+)$. From the equations it follows that the verification ratio determines the direction of the effect of verification bias.

If the test results are verified randomly with respect to the test result the verification ratio VR = 1. In this case the corrected $P(T^+|D^+)$ and $P(T^+|D^-)$ will equal the apparent $P(T^+|D^+,V^+)$ and $P(T^+|D^-,V^+)$.

In diagnostic radiology verification bias usually results in positive test results being verified whereas negative test results tend to be unverified. In the above example patients who had ultrasound studies consistent with appendiceal disease were operated on whereas patients with a negative ultrasound tended to be unverified. The verification ratio VR is thus larger than 1 for many radiological diagnostic tests. From the equations we infer that the corrected $P(T^+|D^+)$ and $P(T^+|D^-)$ will be smaller than the apparent $P(T^+|D^+,V^+)$ and $P(T^+|D^-,V^+)$. (For VR = 1 we derived $P(T^+|D^+) = P(T^+|D^+,V^+)$, thus for VR > 1 we derive $P(T^+|D^+) <$ the $P(T^+|D^+,V^+)$. Likewise, $P(T^+|D^-) < P(T^+|D^-,V^+)$.) Intuitively this makes sense, because if mainly negative test results are unverified, we miss false negatives and therefore $P(T^+|D^+,V^+)$ is an overestimate of $P(T^+|D^+)$. At the same time we miss true negatives thereby underestimating specificity, which is equivalent to overestimating $P(T^+|D^-)$. Correcting for verification bias will thus result in lower $P(T^+|D^+)$ and $P(T^+|D^-)$ than those calculated from the group of verified patients, if the probability of verification is larger in those with a positive test result than in those with a negative test result.

An example of a diagnostic test in which the verification ratio VR < 1 is the L/S ratio performed to test for fetal pulmonary maturity. The gold standard for this diagnostic test is the development of respiratory distress syndrome in the newborn. A L/S ratio may be considered verified if the baby is delivered within a short time after performing the test. Waiting longer would mean that the lungs can mature further, implying that the test result no longer represents the lung maturity at birth. If the test result is positive (ie. suggesting immaturity of the fetal lungs) the obstetrician will do his/her best to delay delivery and the test result will be unverified. Verification of the diagnosis "fetal pulmonary immaturity" will thus tend to be greater among negative (mature) test results than among positive (immature) tests.

A radiological example of a diagnostic test in which the verification ratio VR < 1 is CT evaluation for invasive malignant disease of eg. the cervix or ovaries. If the test result is positive (ie. the test shows evidence of invasive disease) the patient will be treated with radiotherapy and/or chemotherapy, and operative verification will not necessarily be performed. Thus verification of the diagnosis "invasive malignant disease" will tend to be greater among negative test results than among positive results.

From the derived equations we deduce that if VR < 1 the corrected true and false positive rates are larger than the apparent rates. (For VR = 1 we derived $P(T^+|D^+) = P(T^+|D^+,V^+)$, thus for VR < 1 we derive $P(T^+|D^+) > P(T^+|D^+,V^+)$. Likewise $P(T^+|D^-) > P(T^+|D^-,V^+)$.)

6. Implications of correcting for verification bias

If verification depends on test outcome, correcting for this bias will change the sensitivity and specificity of any particular cutoff value of the test variable. However, correcting the data does not necessarily change the ROC curve to a large extent. If the uncorrected and corrected curves are similar, verification bias shifts the operating points along the ROC curve, but does not effect overall test performance. In such cases, although individual sensitivities and specificities can

be seriously biased, the area under the ROC curve is a measure of performance which is less sensitive to bias. Correcting the data set for verification bias implies adjusting the chosen cutoff value of the test variable to a different criterion. This means that if we were to construct a utility structure, the optimal operating point could be seriously affected by verification bias.

IX. Uninterpretability bias

Uninterpretability bias may arise when not all subjects or specimens tested provide interpretable test results. Uninterpretability occurs, for example, when bowel gas interferes with an ultrasound examination; due to bowel movement while performing a digital subtraction angiography; or an unsuccessful transhepatic cholangiography because the bile tract is undilated. By simply ignoring uninterpretable tests the calculated test characteristics may be biased (4,3). This is especially important when comparing diagnostic tests.

An approach to this possible bias is to consider uninterpretable tests as an additional test result (3). Using this approach one has to consider the repeatability of the test and whether or not the uninterpretability is related to true disease status. If the reason for uninterpretability is intrinsic, then repeating the test would again result in an uninterpretable test result. If it is also random with respect to disease, then the estimated test parameters are unbiased. Examples of such situations are: claustrophobia in a patient undergoing an MRI or an ultrasound of the pancreas which is uninterpretable because of obesity of the patient.

Situations in which uninterpretability is intrinsic, but possibly correlated with true disease status, are different. In the example of ultrasound for appendiceal disease, the test is uninterpretable if the appendix is not visualized (17). Repeating the ultrasound would again result in an uninterpretable test. However, the fact that the appendix cannot be visualized, in itself, contains diagnostic information because a healthy appendix is very thin and therefore difficult to visualize, as opposed to a swollen inflamed appendix.

By calculating the interval likelihood ratio (interval LR) of the uninterpretable results one can determine if uninterpretability is associated with disease or not. The interval likelihood ratio of uninterpretable test results I^- is (see section I.):

$$intervalLR_{I^-} = \frac{P(I^- \mid D^+)}{P(I^- \mid D^-)}$$

The original results of the study on ultrasound examination for appendiceal disease, are given in table 5. For a positive US result the interval LR equals infinity, for a dubious test result the interval LR equals 0.60 and for uninterpretable results it equals 0.21. This implies that the ranking of the test results according to decreasing interval LR is as follows: positive US, dubious US, uninterpretable US. In other words, an uninterpretable test result (ie. a non-visualized appendix)

is most likely to occur in the absence of appendiceal disease, and can therefore be used as a negative test result. Without calculating interval likelihood ratio's, the authors of the paper have come to the same conclusion based on pathophysiological considerations.

Table 5. Ultrasound for appendiceal disease.

	US positive	US dubious	appendix not visualized
appendiceal disease	39	3	10
no appendiceal disease	0	3	28
$P(R_i\|D^+)$ $P(R_i\|D^-)$	0.75 0.00	0.06 0.10	0.19 0.90
interval LR	infinity	0.60	0.21

Constructing ROC curves involves more test results than in the given example. However, the method is identical. After calculating the interval LR's of all test results and uninterpretable results, we rank the uninterpretable results among the other test results, according to decreasing interval LR. Subsequently, we include the uninterpretable results into the ROC analysis as an additional test result.

X. REFERENCES

1. Bamber D: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psych 1975; 12: 387-415.

2. Begg CB, Greenes A. Assessment of Diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 207-215.

3. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. J Chron Dis 1986; 39: 575-584.

4. Begg CB, McNeil BJ. Assessment of Radiologic tests: control of bias and other design considerations. Radiology 1988; 167: 565-569.

5. Centor RM, Schwartz JS: An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. Med Decis Making 1985; 5: 149-156.

6. Centor RM: A Visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. Med Decis Making 1985; 5: 139-148.

7. Dorfman DD and Alf E. Maximum Likelihood Estimation of parameters of signal detection theory - a direct solution. Psychometrika 1968; V 33 no 1: 117-124.

8. Egan JP. Signal Detection Theory and ROC Analysis. Academic Press 1975.

9. England WL. An exponential model used for optimal threshold selection on ROC curves. Med Decis Making 1988; 8/2: 120-131.

10. Gray R, Begg CB, Greenes RA. Construction of Receiver Operating Characteristic Curves when Disease Verification is subject to selection bias. Med Decis Making 1984; 4: 151-164.

11. Green DM and Swets JA. Signal Detection and Psychophysics. John Wiley and Sons 1966.

12. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148; 839-843.

13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29-36.

14. Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. Med Decis Making 1988; 8: 197-203.

15. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24: 234-245.

16. Metz CE: ROC methodology in radiologic imaging. Invest Radiol 1986; 21: 720-733.

17. Puylaert JBCM, Rutgers PH, Lalisang RI, de Vries BC, vd Werf SDJ, Dorr JPJ, Blok APR. A prospective study of ultrasonography in the diagnosis of appendicitis. N Engl J Med 1987; 317: 666-669.

18. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299: 926-930.

19. Rifkin RD. Maximum Shannon Information content of diagnostic medical testing. Med Decis Making 1985; 5: 179-190.

20. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Butterworth Publishers, Boston, 1988.

21. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 1986; 99: 181-198.

22. Swets JA and Pickett RM. Evaluation of Diagnostic Systems. Academic Press 1982.

Chapter VI

CT AND MRI ASSESSMENT OF ENT TUMORS USING ROC METHODOLOGY[1]

I. INTRODUCTION
II. MATERIAL AND METHODS
1. The study population
2. CT and MRI technique
3. Compartments of tumor extension
4. Verification of tumor extension
5. ROC analysis
III. RESULTS
1. Results of the signal analysis
2. Tumor extension and results of the ROC analysis
IV. DISCUSSION
V. REFERENCES

ABSTRACT

Neoplastic disease of the nasopharynx, the paranasal sinuses and the parapharyngeal space requires thorough assessment of location and extension in order to plan appropriate treatment. This study evaluates computer tomography (CT) and magnetic resonance imaging (MRI) in the workup of malignant and non-malignant tumors of the nasopharynx, the paranasal sinuses and the parapharyngeal space in 76 patients. An attempt is made to characterize histopathology on MRI by analysing the signal intensities on T1- and T2-weighted images relative to muscle and brain tissue. The test performance of CT and MRI in the assessment of tumor extension are compared with ROC methodology.

Although no definitive conclusions can be made as to the histopathology on the basis of the signal intensities on MRI, some tumors show characteristic images. ROC analysis of the performance of CT and MRI in the assessment of extension of neoplastic disease of the nasopharynx, the paranasal sinuses and the parapharyngeal space, demonstrates no statistical significant

---

[1]Co-authors: Ragnhild GM de Slegte, Geerten J Gerritsen, Henk Speelman; Dept of Diagnostic Radiology and Dept of Otolaryngology & Head and Neck Surgery, Free University Hospital, Amsterdam.

Submitted for publication.

difference in overall test performance. However, in evaluating regions involving predominantly soft tissue structures and comparatively large bony structures MRI is superior to CT, while in evaluating regions involving thin bony structures, CT performs better than MRI.

I. INTRODUCTION

Computer tomography (CT) has an established role in assessing benign and malignant lesions of nose and paranasal sinuses, the nasopharynx and the parapharyngeal space (3). Magnetic resonance imaging (MRI) has in this respect possible advantages over CT (11,15), however, these have as yet not been unambiguously demonstrated. The potential advantages of MRI over CT in the assessment of ENT tumors are the possibility for three plane imaging, less artifacts in the region of the skull base and better delineation of soft tissues. However, MRI also has drawbacks: in the evaluation of thin bony structures and in the identification of calcification, CT seems better than MRI.

The most common malignant neoplasms of nose and paranasal sinuses, the nasopharynx and the parapharyngeal space are squamous cell carcinomas and adenocarcinomas (15). Malignancies of the paranasal sinus, nasopharynx and parapharyngeal space grow invasively and on admission patients usually present with advanced disease. The diagnostic workup requires a thorough assessment of tumor location and extension into adjacent areas in order to plan appropriate treatment. Most patients seen at our hospital present with advanced disease for which the treatment of choice is surgery with post-operative irradiation. However the choice between radical surgery, radiotherapy, chemotherapy or a combination of these depends on histopathological findings and staging.

Benign neoplastic disease of nose and paranasal sinuses, parapharyngeal space and nasopharynx include such lesions as poliposis and inverting papilloma. These lesions can be extensive and, as with malignant neoplasms, the extent of disease has to be assessed before proceeding to surgery. Glomus tumors, another benign entity, often occur in multiple locations requiring assessment before proceeding to treatment.

This paper presents the results of a comparative study of CT and MRI in the preoperative workup of patients with space-occupying lesions of nose and paranasal sinuses, the nasopharynx and the parapharyngeal space. The goals of the study are twofold: tumor characterization on MRI and comparison of CT and MRI test performance in assessing tumor extension. The signal intensities on T1- and T2-weighted MR images are analysed to determine if particular tumors can be characterized by their relative signal intensities. We compare the signal intensities relative to muscle and relative to brain. Tumor extension is graded according to the UICC classification system (International Union against Cancer)(10), modified to suit treatment protocols in our hospital. The performance of CT and MRI in the assessment of tumor extension are compared with receiver operating characteristic (ROC) methodology.

II. MATERIAL AND METHODS

1. The study population

From 1985 to 1988 seventy-six patients with suspected malignant or benign lesions of nose and paranasal sinuses, the nasopharynx and the parapharyngeal space were studied. The study population consisted of 44 males and 32 females, their age varying from 4 months to 88 years. Informed consent was given orally.

All tumors, except the glomus tumors, were biopsied to determine the nature of the neoplasm and to compare the true histopathology with the pattern of the signal intensities on MRI T1- and T2-weighted images. The definitive diagnosis in suspected glomus tumors was made by means of angiography. Table 1 gives the histopathology of the cases and the frequency of each.

Table 1. Tumor histopathology and frequency of each.

MALIGNANT DISEASE		NON MALIGNANT DISEASE	
squamous cell carcinoma	17	glomus tumor	5
adenocarcinoma	9	poliposis	4
adenocystic carcinoma rhabdomyosar-	6	inflammation	2
coma	4	encephalocele	2
melanoma	3	osteoma	1
chondrosarcoma	2	papilloma inversum	2
plasmocytoma	2	osteomyelitis	2
malignant lymphoma	2	pleomorphic adenoma	2
fibrosarcoma	1	cyst/mucocele	2
fibrous histiocytoma	1	lipoma	1
ameloblastoma	1	nasal glioma	1
		haemangioma	1
		angiofibroma	1
		granuloma	1
		giant cell tumor	1
TOTAL MALIGNANT	48	TOTAL NON MALIGNANT	28

2. CT and MRI technique

CT scanning and MR imaging were performed to evaluate the extent of the lesions. 73 of the 76 patients underwent both CT and MR imaging. Two patients refused the CT examination and in one patient MR imaging failed due to claustrophobia. These 3 patients were excluded from the comparative analysis.

The CT examinations were performed on a Philips 350 Tomoscan (high resolution mode). The images were made in transverse and coronal directions with a slice thickness of 3 mm and a slice-increment of 6 mm. If not enough information was obtained additional 3 mm consecutive slices were made.

The MR images were made on a 0.6 Tesla Technicare superconductive system in transverse, coronal and sagittal directions with a slice thickness of 1 cm before December 1986 and a slice thickness of 0.4 to 0.75 cm since 1987.

The pulse sequences used are:

- T1-weighted images :	TR $\leq$ 600 msec
	TE $\leq$ 38 msec
- T2-weighted images :	TR $\geq$ 1000 msec
	TE $\geq$ 64 msec
- late T2-weighted images :	TR $\geq$ 2000 msec
	TE more than 4 echoes

CT scanning was usually performed before the MRI study. The CT and MR images were evaluated in the clinical setting by a panel of 3 readers, the reported diagnosis being one of majority rule. The images were read without prior knowledge of the final diagnosis or the result of the other diagnostic test.

3. Compartments of tumor extension

To define the location of tumor extension we used the classification system of the International Union against Cancer (UICC) (10), modified to suit treatment protocols in our hospital. The classification we used applies to tumors considered to arise within the maxillary sinus. However, tumors originating at the alveolar ridge or upper jaw and tumors in the ethmoid region, although not strictly part of the UICC classification system, can be staged in a similar fashion.

In the UICC classification (10):

- stage T1 tumors are confined to the antral mucosa;
- stage T2 tumors show erosion or destruction of the infrastructure including the hard palate and/or the middle nasal meatus;
- stage T3 tumors invade the cheek, the posterior wall of the maxillary sinus, the floor or medial

wall of the orbit and/or the anterior ethmoid sinus;
- stage T4 tumors extend to the orbital contents and/or the cribriform plate, the posterior ethmoid or sphenoid sinuses, the nasopharynx, the soft palate, the pterygomaxillary or temporal fossae and/or the base of the skull.

Based on the UICC classification system we defined 5 compartments to which tumor can extend as follows:

Compartment 1: Involvement of the maxilla or soft tissue of the cheek, (corresponds with stage T2 or stage T3 according to UICC staging)
Compartment 2: Involvement of the medial wall of the maxillary sinus with tumor extension into the nasal cavity or palate, or primary tumor of the nasal cavity (corresponds with stage T2)
Compartment 3: Involvement of the anterior ethmoid cells and/or orbital involvement or primary tumor localized in ethmoidal or frontal sinuses (corresponds with stage T3)
Compartment 4: Involvement of pterygopalatine fossa, infratemporal fossa or nasopharynx (corresponds with stage T4)
Compartment 5: Involvement of the middle cranial fossa, cribriform plate, clivus or skull base (corresponds with stage T4)

Divisions into the listed compartments has practical consequences for the treatment choice in our hospital. In patients with tumor extension to compartments 1, 2 or 3 radical surgery is attempted. Patients with tumor extension to compartments 4 and 5 are treated with a debulking procedure, combined with radiotherapy and/or chemotherapy. Tumors originating in compartment 4, that is in the nasopharynx, are irradiated only. Contra-indications for radical surgery are distant metastasis and/or extension to compartments 4 and/or 5.

The location and extent of benign lesions were classified according to the same 5 compartments.

4. Verification of tumor extension

Tumor extension was, whenever possible, verified by histology of resected tissue obtained at surgery or obtained through endoscopy, together with macroscopic findings. The gold standard for compartments 1 to 4 was histology as described. The gold standard for compartment 5 was histology, if obtained, or unambiguous neurological signs related to the cranial nerves, if present. Glomus tumors were verified by means of angiography. In 56 of the 76 patients the presence or absence of tumor could be verified for all 5 compartments. In the remaining cases only verification for 2, 3 or 4 compartments was possible. Considering extension to each compartment as a separate case a total of 285 diagnoses made on CT and MRI could be compared to the gold standard.

5. ROC analysis

Receiver Operating Characteristic (ROC) methodology is used to compare CT and MRI test performance in the assessment of tumor extension.

A ROC curve plots pairs of true positive and false positive rates at various confidence thresholds for the diagnosis. The area under the ROC curve is a measure of diagnostic system performance, a useful parameter for comparing diagnostic tests (12).

A likelihood was assigned for the diagnosis "extension of tumor" to each compartment, using a five point rating scale (Table 2):

1. definitively, or almost definitively, negative (no extension),
2. probably negative (extension unlikely),
3. equivocal (extension possible),
4. probably positive (extension likely),
5. definitively, or almost definitively, positive (extension present).

Table 2. Cross tabulation of CT and MRI results.

(a) CT DIAGNOSIS

	positive	probably positive	equi-vocal	probably negative	negative
FINAL DIAGNOSIS					
positive	89	11	4	4	6
negative	4	6	2	6	143
UNVERIFIED	26	4	9	4	32

(b) MRI DIAGNOSIS

	positive	probably positive	equi-vocal	probably negative	negative
FINAL DIAGNOSIS					
positive	75	15	9	10	5
negative	3	3	7	10	138
UNVERIFIED	23	5	15	5	27

To construct the ROC curve the true positive and false positive rates are calculated at all possible cutoff values on the likelihood scale. In this study the area under the ROC curve and the standard error of the area are calculated using the Mann-Whitney U statistic. The Mann Whitney method is a convenient method and has the advantage of being non-parametric (4,5,8), meaning that no assumptions are made as to the underlying distribution of test results. The area under the ROC curve is underestimated with the Mann-Whitney method compared to the classically used maximum likelihood estimation method (4,6). However, if comparison of two curves is the point of interest this is of little consequence.

CT and MRI test performance in assessing tumor extension are first compared for all the data pooled in one analysis and then for benign and malignant lesions separately. Subsequently, the tests are compared for the five mentioned compartments separately and, furthermore, for various compartments pooled. Subdividing the study population for compartment and malignancy is precluded because of sparseness of the data.

We use a one-sided paired t-test to test for a difference in area under the ROC curves of CT and MRI, taking into account the correlation between test results derived from the same cases (7). Taking the correlation between test results into account makes the statistical test more powerful in detecting a difference, if one is present. In using a one-sided, as opposed to a two-sided, t-test we assume, on theoretical grounds, that one test will perform equal or better than the other test. More specifically, for the combined data and for soft tissue or large bony structures, we assume that MRI will perform equal or better than CT, but for fine bony structures we assume that CT will perform equal or better than MRI. If this assumption is not considered valid the given p-values should be multiplied by 2.

III. RESULTS

1. Results of the signal analysis

We attempt to characterize the histopathology of the tumors on MRI by analysing the signal intensities of the tumors comparative to muscle and brain on T1-, T2- and late T2-weighted images. Distinguishing benign from malignant lesions on T1-weighted images (Table 3) is impossible on the basis of the signal intensity of the tumor. On T2-weighted images (Table 3) more than half of the malignant tumors have signal intensities hyperintense relative to muscle and isointense relative to brain while less than a quarter of the benign lesions have the same signal intensity pattern on T2-weighted images. On late T2-weighted images signal intensities hyperintense relative to muscle and isointense relative to brain also occur more frequently in malignant lesions. On late T2-weighted images benign lesions have either an isointense or white signal relative to muscle.

Table 3. Signal intensities on MR images of malignant and non malignant lesions. Table of the relative frequency (in percentages) that a signal intensity occurs on T1-, T2- and late T2-weighted images among malignant and non-malignant lesions. For example, of all malignant lesions 64.3% have a hyperintense signal compared to muscle and an isointense signal compared to brain on T2-weighted images.

image	signal relative to muscle*	signal relative to brain*	frequency (%) among malignant lesions	frequency (%) among non malignant lesions
T1-weighted	2	4	2.6	
	3	2	33.3	12.5
	3	3		8.3
	3	4	2.6	
	4	1		4.2
	4	2	10.3	8.3
	4	3	46.2	50.0
	4	4	2.6	12.5
	5	4	2.6	
	5	5		4.2
T2-weighted	3	2	4.8	
	3	3	2.4	4.8
	4	2		4.8
	4	3	64.3	19.0
	4	4	14.3	23.8
	4	5		4.8
	5	3	4.8	9.5
	5	4	9.5	14.3
	5	5		19.0
late T2-weighted	3	3		11.1
	3	4		11.1
	4	2	11.1	
	4	3	33.3	
	4	4	11.1	
	5	2	3.7	
	5	3	7.4	
	5	4	25.9	33.3
	5	5	7.4	44.4

*1=black 2=hypointense 3=isointense 4=hyperintense 5=white

Table 4. Signal intensities on MR images for squamous cell carcinoma, adeno- carcinoma and adenocystic carcinoma. Table of the relative frequency (in percentages) that a signal intensity occurs on T1-, T2- and late T2-weighted images among squamous-, adeno- and adenocystic carcinomas. For example, of all squamous cell carcinomas 64.3% have a hyperintense signal compared to muscle and an isointense signal compared to brain on T1-weighted images.

image	signal relative to muscle*	signal relative to brain*	frequency (%) among squamous carcinomas	frequency (%) among adeno carcinomas	frequency (%) among adenocystic carcinomas
T1-weighted	2	4	7.1		
	3	2	14.3	62.5	60.0
	3	4	7.1		
	4	2	7.1	25.0	
	4	3	64.3	12.5	40.0
T2-weighted	3	2		11.1	
	4	3	66.7	44.4	100.0
	4	4	33.3	11.1	
	5	3		11.1	
	5	4		22.2	
late T2-weighted	4	2	10.0		33.3
	4	3	50.0	28.6	
	4	4	10.0	14.3	
	5	2		14.3	
	5	3	10.0	14.3	
	5	4	20.0	14.3	66.7
	5	5		14.3	

*1=black 2=hypointense 3=isointense 4=hyperintense 5=white

The squamous cell carcinomas are usually hyperintense relative to muscle and isointense relative to brain on T1-, T2- and late T2-weighted images (Table 4). Adenocarcinoma (Table 4) are often isointense relative to muscle and hypointense relative to brain on T1-weighted images, but have a wide variety of signals on T2- and late T2-weighted images. Adenocystic carcinomas (Table 4) are often isointense relative to muscle and hypointense relative to brain on T1-weighted images. On T2-weighted images these carcinomas show hyperintense signal relative to muscle and isointense signal relative to brain. Furthermore, these lesions often show mixed signal intensities and the mucous produced by the tumors can be distinguished on the late T2-weighted images by its hyperintense or white signal.

In general, mucous producing disease and retained secretions, including benign lesions such as mucoceles and sinusitis, can be distinguished because mucous has a hyperintense or white signal intensity on T2- and late T2-weighted images. Glomus tumors show a characteristic image due to the flow void phenomenon of flowing blood which results in a lack of signal in the vessel lumen. With MR imaging, tumors primarily originating from the soft tissues, such as fibrosarcomas, rhabdomyosarcomas and lymphomas can be distinguished from normal soft tissue structures by their hyperintense signal relative to muscle on T2- and late T2-weighted images. About the remaining tumors in our study no conclusions can be made as to the signal intensities because the subgroups are too small.

2. Tumor extension and results of the ROC analysis

Table 5. Results of the ROC analysis. Tabulated are the number of positive and negative cases contributing to the curve, area under the ROC curve and the standard error, for all the data combined, stratified for benign versus malignant disease, stratified by compartment and for various compartments combined. The p-values given are for one-sided paired t tests.

ROC curve	positive	negative	CT: area	CT: error	MRI: area	MRI: error	p
all data combined	114	161	0.95	0.02	0.95	0.01	ns
malignant disease	81	84	0.94	0.02	0.93	0.02	ns
benign disease	33	77	0.97	0.03	0.98	0.02	ns
compartment:							
1) maxilla/ cheek	21	33	0.94	0.04	0.96	0.03	0.25
2) medial wall maxillary sinus/ nasal cavity/ palate	34	21	0.97	0.03	0.94	0.03	0.18
3) orbit/ ethmoid/ frontal	24	30	0.92	0.05	0.94	0.03	0.21
4) pterygopalatine/ infratemporal/ nasopharynx	18	36	0.95	0.03	0.95	0.03	ns
5) skull base/ middle cranial fossa/ cribiform/ clivus	17	41	0.99	0.01	0.98	0.01	0.21
compartments combined:							
1 & 3	45	63	0.93	0.03	0.96	0.02	0.14
2 & 5	51	62	0.97	0.02	0.95	0.02	0.17

Of the verified cases the percentage that actually had extension of tumor growth to the five compartments is as follows:

- Compartment 1: the maxilla or soft tissue of the cheek: 39%
- Compartment 2: the medial wall of the maxillary sinus, nasal cavity or palate: 62%
- Compartment 3: the orbital involvement, ethmoidal or frontal sinuses: 44%
- Compartment 4: the pterygopalatine and infratemporal fossa or nasopharynx: 33%.
- Compartment 5: the middle cranial fossa, cribriform plate, clivus or skull base: 29%

In three cases extension of tumor on both CT and MRI was reported as definitively positive (Table 2) while at surgery thickened mucosa as a result of inflammation was found. In one case CT was reported definitively positive for extension to compartment 1, MRI being equivocal, while no extension was found. This was in a patient with chondrosarcoma of the nasal cavity. Extension was falsely reported as definitively negative in 6 cases on CT and 5 cases on MRI (Table 2), the pathology in these cases being adenocarcinoma, squamous cell carcinoma or poliposis. There was no clear pattern identifying the cause of error in the false negative cases.

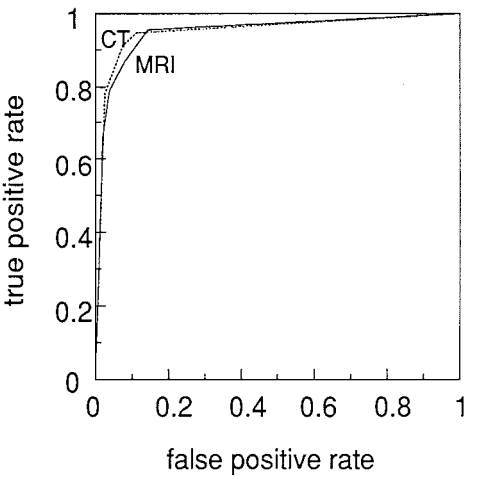


Figure 1. ROC curves of CT (dashed line) and MRI (solid line) in the evaluation of ENT tumor extension.

The ROC curves of CT and MRI in assessing tumor extension are presented in figure 1. The ROC curves shown are for the five compartments combined. No difference between CT and MRI test performance can be demonstrated, the area under the ROC curves being 0.95 for both CT and MRI (Table 5) with standard errors of 0.02 for CT and 0.01 for MRI. The ROC analysis

was done separately for benign and malignant tumors (Table 5). For neither benign nor malignant lesions can a statistically significant difference between CT and MRI be demonstrated. Both modalities perform better for benign disease than for malignant disease. The ROC analysis was also done for the 5 compartments separately (Table 5). Comparing CT and MRI for each compartment we found that CT performs better than MRI in evaluating the medial wall of the maxillary sinus, nasal cavity and palate, but the significance level is only 0.18 (one-sided). The lack of statistical significant differences is partly due to sparseness of the data. For this reason we pooled the data for compartments 1 and 3, areas involving predominantly soft tissue and relatively large bony structures, and for compartments 2 and 5, which involve many thin bony structures. For compartments 1 and 3 (that is the maxilla, cheek, orbit, ethmoid and frontal sinuses) MRI performs better than CT at a significance level of 0.14 (one-sided) (Table 5). However, for the compartments 2 and 5 (that is the medial wall of the maxillary sinus, the nasal cavity, palate, middle cranial fossa, cribriform plate, clivus and skull base) CT performs better than MRI, but only at a significance level of 0.17 (one-sided).

IV. DISCUSSION

In an attempt to characterize the histopathology of ENT tumors on MRI we analysed the signal intensity of the tumors on T1- and T2-weighted images. We compared the signal intensity relative to muscle and brain tissue. Although no definitive conclusion can be made as to tumor histology on the basis of the MRI signal, some tumors show characteristic images. Squamous and adenocystic carcinoma tend to be hyperintense relative to muscle on T2-weighted images and either hyperintense or white (relative to muscle) on late T2-weighted images. As has been reported elsewhere (11) we found that mucous producing lesions and retained secretions have high signal (hyperintense or white) on T2- and late T2-weighted images. Glomus tumors are distinguished by the flow void phenomena. In general, the anatomy of soft tissue structures is easier to distinguish on MRI than on CT.

The study presented was done within the clinical setting which implies that internal noise was not controlled for. Day-to-day variations in reader performance and variations in the performance of the panel of readers could possibly have degraded the test performance and/or masked differences between the two tests evaluated. However, we have chosen to read the images in the daily routine so as to evaluate test performance in the clinical context, with its variations (9). Between reader variance was not considered (9,14) because all images were read by the same panel of readers. It is possible that the performance of the tests is inflated because of panel judgement, however, this is not likely to have a large effect. Furthermore, it is not uncommon in clinical practise that images are discussed and reported by a team of radiologists. All things considered our conclusions are somewhat limited in that they are, strictly speaking, only applicable to the readers studied. However, we feel that our readers are representative for readers in many academic centers and that our general conclusions hold for other settings than our own.

Extension of tumor was not verified for all compartments in all patients which made us consider the potential problem of verification bias (1,2). However, the unverified cases are distributed over all test results (Table 2) and therefore selection for verification is not biased in this analysis, implying that the unverified cases can be disregarded without affecting the results.

In the assessment of tumor extension most false positive and false negative reports occurred simultaneously in CT and MRI examinations, suggesting that the errors of one test will not be captured by the other test. False positive results were mainly caused by thickened mucosa which was mistaken for tumor extension. False negative results did not show any clear pattern identifying the cause of error.

The areas under the ROC curves of CT and MRI in evaluating tumor extension show no statistically significant differences for all the data combined, for most of the compartments separately and for malignant and benign disease separately. In analysing the compartments independently, CT performs better than MRI in evaluating the medial wall of the maxillary sinus, the nasal cavity and palate, but the significance level is low. Because our sample size is somewhat limited, especially when subdivided by compartment, a small difference in area can go undetected (7). Pooling for the five compartments creates a larger sample size, making the statistical test more powerful, but has the disadvantage of masking differences because of an averaging effect. For this reason we pooled the data for various subsets of the five compartments, pooling compartments for which there is a theoretical basis that one test will perform better than the other. Analysing the pooled subsets we found that MRI performs better than CT in evaluating the maxilla, cheek, orbit, ethmoid and frontal sinuses, which involve predominantly soft tissue structures and comparatively large bony structures. However, in evaluating the medial wall of the maxillary sinus, the nasal cavity, palate, skull base, middle cranial fossa, cribriform plate and clivus, regions involving many thin bony structures, CT performs better than MRI.

Summarizing, although we cannot prove it formally due to limited sample size, CT and MRI do not seem to differ in overall test performance in staging ENT tumors of the nasopharynx, paranasal sinuses and parapharyngeal space. However, in evaluating regions involving predominantly soft tissue structures and comparatively large bony structures MRI is superior to CT, while in evaluating regions involving many thin bony structures, CT performs better than MRI.

V. REFERENCES

1. Begg CB, Greenes A. Assessment of Diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 207-215.

2. Begg CB, McNeil BJ. Assessment of Radiologic tests: control of bias and other design considerations. Rad 1988; 167: 565-569.

3. Carter BL. Upper Aerodigestive Tract and Neck. In: Haaga JR, Alfidi RJ. Computer Tomography of the Whole Body. Vol I: 445-478. 2ed. Mosby Co., St. Louis, 1988.

4. Centor RM, Schwartz JS. An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. Med Decis Making 1985; 5: 149-156.

5. Centor RM. A Visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. Med Decis Making 1985; 5: 139-148.

6. Dorfman DD and Alf E. Maximum Likelihood Estimation of parameters of signal detection theory - a direct solution. Psychometrika 1968; V 33 no 1: 117-124.

7. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Rad 1983; 148; 839-843.

8. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Rad 1982; 143: 29-36.

9. Hanley JA. Alternative Approaches to Receiver Operating Characteristic Analyses. Rad 1988; 168: 568-570.

10. Hermanek P, Sobin LH. TNM Classification of Malignant Tumours. UICC International Union against Cancer. 4ed, 1987, Springer-Verlag, Berlin.

11. Lloyd GAS, Lund VJ, Phelps PD, Howard DJ. Magnetic resonance imaging in the evaluation of nose and paranasal sinus disease. Br J Radiology 1987; 60: 957-968.

12. Metz CE. ROC methodology in Radiologic Imaging. Invest Radiol 1986; 21: 720-733.

13. Slegte RGM de, Gerritsen GJ, Nauta JJ, Hoen MB, Crezee FC. Comparative study of MRI versus CT for the diagnostic workup of lesions in the nose and paranasal sinuses. In: Theo HM Falke, ed. Essentials of Clinical MRI. Martinus Nijhoff, 1988, 1ed.

14. Swets JA and Pickett RM. Evaluation of Diagnostic Systems. Academic Press 1982.

15. Ziedes des Plantes BG, de Slegte RGM, Gerritsen GJ, Sperber M, Valk J, Kaiser MC, Crezee FC. Malignant Lesions of the Paranasal Sinuses. In: Partain CL, Price RR, Patton JA, Kulkarni MV, James AE, eds. Magnetic Resonance Imaging. Philadelphia: Saunders Co., 1988, 2ed. Vol. I: 308-321.

Chapter VII

ROC ANALYSIS OF THE CLINICAL, CT AND MRI DIAGNOSIS OF ORBITAL SPACE-OCCUPYING LESIONS[1]

I. INTRODUCTION
II. MATERIAL and METHODS
1. The study population and diagnostic procedures
2. ROC methodology
3. Correction for verification bias
III. RESULTS
IV. DISCUSSION
V. ACKNOWLEDGMENT
VI. REFERENCES

ABSTRACT

A study was done comparing clinical evaluation, computer tomography (CT) and magnetic resonance imaging (MRI) performed for suspected orbital space-occupying lesions (SOL). Some illustrative cases are presented. Receiver operating characteristic (ROC) curves are used to compare the diagnostic tests. The methodological issue of verification bias is addressed and the constructed ROC curves are adjusted for biased selection for verification. The test performance of CT and MRI do not differ significantly from that of clinical evaluation. Correcting for verification bias MRI shows an advantage over CT, however the statistical significance is low (p value=0.15, one-sided test).

In conclusion, although MRI performs better than CT, neither MRI nor CT can be demonstrated to provide significantly more information than clinical evaluation in the diagnosis of orbital SOL.

---

[1]Co-authors: Ragnhild GM de Slegte and Marion F Hoogesteger, Department of Diagnostic Radiology and Department of Ophthalmology, Free University Hospital, Amsterdam.

Reproduced with permission from the journal ORBIT, in press 1989; 8/3: 173-187.

I. INTRODUCTION

Although magnetic resonance imaging (MRI) produces beautiful images of orbital anatomy and pathology, the clinical significance of MRI for orbital space-occupying lesions (SOL) is as yet not clear. In previous papers on MRI of orbital SOL attention has been paid to characterization of lesions on the basis of signal intensity on T1- and T2-weighted images. A number of lesions appear to have characteristic combinations of signal intensities (4,13,14).

Ocular melanotic melanoma, a common malignant orbital tumor in adults, has an intermediate signal on T1-weighted images and a low signal on T2-weighted images (4,7). This characteristic combination of signal intensities is due to the paramagnetic properties of free radicals in melanin which shorten T1 and T2 relaxation times. Retinoblastoma, which occurs at a very young age, has an intermediate signal on T1- and a low signal on T2-weighted images. Dermoids have a high signal intensity on both T1- and T2-weighted images.

In retrobulbar lesions the main advantages of MRI over CT are the oblique view parallel to the optic nerve and demonstration of the intracanalicular portion of the nerve. As yet characterization of optic nerve tumors is not possible (1). Optic nerve glioma and optic nerve sheath meningioma both have low signal intensity on T1- and T2-weighted images.

Although numerous reports have been published describing MRI images of orbital SOL, no attempt has as yet been made to actually measure MRI test performance in comparison to other diagnostic methods. This paper presents a pilot study in which the performance of clinical assessment, CT and MRI in the evaluation of malignant SOL of the orbit are compared using receiver operating characteristic (ROC) methodology. Instead of comparing only imaging techniques as is often done we assess the incremental value of CT and MRI over and above the information provided by clinical evaluation.

II. MATERIAL and METHODS

1. The study population and diagnostic procedures

All patients clinically suspected of having orbital SOL who underwent clinical evaluation, CT and MRI in the years 1985, 1986 and 1987 were included in the study. The test results were reviewed retrospectively by at least two of the authors, taking into account the written reports. Informed consent was given orally.

Clinical evaluation consisted of taking the history, slit lamp examination of the eye and ophthalmoscopy.

Table 1. DIAGNOSES, FREQUENCIES AND SIGNAL INTENSITIES. Definitive diagnoses made and the frequency of each diagnosis. Signal intensities of the lesions on T1- and T2-weighted MRI images are given where relevant.

DEFINITIVE DIAGNOSIS	FREQUENCY	MRI signal T1 image(*)	MRI signal T2 image(*)
POSITIVE			
melanotic melanoma	9	3/4	2
retinoblastoma	1	3	2
retrobulbar rhabdomyosarcoma	2	2/3	V
carcinoma of the lacrimal gland	2	3	M
invasion of the bony orbita	1		
NEGATIVE			
retinal detachment	8	3/4	3/4
persistent hyperplastic primary vitreous	2	2/3	V
colaboma	1	2	4
retrobulbar pseudotumor	2	2	2
optical nerve neurinoma	1	2/3	5
benign histiocytoma	1	2	1
varicose vein plus thrombus	1	3	4
arteriovenous malformation	1	1	1
carotid fistula	1	1	1
no retrobulbar invasion from malignant ocular tumor	10		
miscellaneous	5		

* MRI signal intensities: 1=black, 2=low signal, 3=intermediate signal, 4=high signal, 5=white, M=mixed signal, V=variable between cases.

The CT examination was done on a Philips 350 Tomoscan before and after intravenous contrast administration. Axial and coronal slices were made using a slice thickness of 1.5 mm and a slice increment of 3 mm.

MRI examination was performed on a 0.6 Tesla Technicare machine in most cases using surface coils. If necessary patients were asked to remove eye cosmetics because cosmetics give artefacts. Axial, coronal and oblique (parallel to the optic nerve) slices were made. The slice thickness was 5 mm and the acquisition matrix 256 x 192. Images were made with spin echo sequences. T1-weighted images were made using a TR varying between 300 and 650 msec and a TE between

24 and 38 msec. T2-weighted images were made using a TR varying between 1000 and 3000 msec and a TE between 30 and 148 msec with multiple echoes. MRI failed in 3 patients because of claustrophobia and/or patient movement.

The diagnostic procedures were always done in the same sequence, namely clinical evaluation followed by CT followed by MRI. Clinical presentation was taken into account in making the CT and MRI diagnosis as this is everyday practice. Our objective was to assess the increment in information after doing CT or MRI as compared to the information provided by clinical evaluation alone.

In total 58 diagnoses were made in 50 patients. The diagnoses made or rejected included malignant orbital lesions (such as malignant melanoma), invasive disease from malignant orbital lesion (such as retrobulbar extension of malignant ocular tumor), and benign orbital lesions (such as retinal detachment and pseudotumor). Thus, in addition to discriminating benign from malignant disease, extension of tumor was also considered a diagnosis.

Of the 58 diagnoses made 48 were verified (see table 1) whilst 10 cases were not verified. In 43 cases the diagnosis was verified histologically. In one case the definitive diagnosis was made at surgery, the diagnosis being a varicose vein. In 2 cases angiography provided the definitive diagnosis: one was a carotid fistula and the other an arteriovenous malformation. In 2 cases complete regression after corticosteroid treatment was considered diagnostic for pseudotumor. In 10 cases the diagnosis was not verified because a malignant SOL was considered so unlikely that doing any further procedures would not have been justified.

2. ROC methodology

Test performance can be characterized by several parameters, sensitivity and specificity being the most commonly used. However, these test parameters depend not only on the capacity of the test to distinguish diseased from non-diseased but also on the confidence threshold that the observer uses to decide whether to label a patient as diseased or not (10). Receiver operating characteristic (ROC) curves provide an alternative method to describe test performance. A ROC curve plots the true positive fraction as a function of the false positive fraction. The area under the ROC curve is a measure of diagnostic system performance and is independent of the ultimately chosen confidence threshold. By comparing the areas under the ROC curves of different diagnostic tests one can compare the performance of these tests (8,9,10).

In deriving a ROC curve the easiest approach is to divide the diagnostic truth into two categories. The choice of the division should be clinically relevant, preferably determining the choice of treatment. In this study the diagnostic truths are:

1/ diagnosis positive:
malignant SOL of the orbit or
retrobulbar extension of malignant ocular tumor or
bony invasion by malignant orbital tumor

Table 2. RANKING ON THE LIKELIHOOD SCALE. Frequency tables of ranking on the likelihood scale versus the final diagnosis for the a) clinical b) CT and c) MRI diagnosis. Both verified and unverified cases are tabulated.

(a) CLINICAL DIAGNOSIS

	positive	probably positive	equivocal	probably negative	negative
FINAL DIAGNOSIS					
positive	1	11	3	0	0
negative	0	1	11	6	15
UNVERIFIED	0	0	5	2	3

(b) CT DIAGNOSIS

	positive	probably positive	equivocal	probably negative	negative
FINAL DIAGNOSIS					
positive	9	2	2	0	2
negative	0	0	5	8	20
UNVERIFIED	0	0	2	1	7

(c) MRI DIAGNOSIS

	positive	probably positive	equivocal	probably negative	negative
FINAL DIAGNOSIS					
positive	9	2	3	1	0
negative	0	2	5	8	18
UNVERIFIED	0	0	1	2	7

versus
2/ diagnosis negative:
benign SOL of the orbit or
retinal detachment or
no extension of malignant tumor

In general the distinction between malignant and benign SOL is clinically relevant in that a malignant lesion always requires enucleation of the orbit whilst that is not the case with benign lesions (1).

A likelihood was assigned for the diagnosis using a five point scale:

1. definitively, or almost definitively, negative
2. probably negative
3. equivocal
4. probably positive
5. definitively, or almost definitively, positive

All patients were thus categorized into 1 of 5 test results (ie. likelihood of diagnosis) for each diagnostic test and into 1 of 2 diagnostic truths (see table 2).

To draw the ROC curve the confidence threshold was shifted several times. At all possible cut-off points on the likelihood scale the true positive rate (TPR) and false positive rate (FPR) were calculated. With a 5 point likelihood scale one can derive 4 pairs of {TPR,FPR} giving 4 points of the ROC curve (10). The points {0,0} and {1,1} are additional points inherent to all ROC curves.

The area under the ROC curve was calculated using the Mann-Whitney U statistic (2). This technique for calculating the area under the ROC curve has several advantages: it is relatively simple to calculate and it is non-parametric, implying that a normal distribution of the study population is not a prerequisite (6). Furthermore it is for practical purposes reliable enough (5).

In comparing the areas under the ROC curves a correction was performed for the fact that the data were derived from the same cases. This correction adjusts the standard error and is based on the correlation coefficients between test results obtained from the same cases (8).

3. Correction for verification bias

The data were corrected for verification bias (3,11,12). Verification bias occurs when only a subset of all tested subjects are selected to undergo the definitive procedure to assess true disease status (11,12). If after workup a malignant SOL is considered unlikely, the patient will not undergo any invasive procedures to verify the diagnosis as this is ethically unacceptable. The result of the CT and MRI exams in combination with the clinical diagnosis will influence the decision to do a definitive procedure to assess disease status and thus the test results will

determine the probability that the diagnosis will be verified or not. Biased selection for verification can have an effect on the calculated test characteristics if we simply ignore the unverified cases. A technique exists to mathematically correct for verification bias (11,12).

The basic assumption for the mathematical correction for verification bias is: the probability of verification is independent of true disease status and the decision to verify depends only on the test result and clinical features. In other words: the influence of true disease status on selection for verification is only through the effect the disease has on the test result and clinical features. This basic assumption is equivalent to saying that the predictive values will not be affected.

To understand how the correction for verification works it is useful to consider a situation where selection for verification depends only on the test under consideration and the test result is dichotomous (ie. either positive or negative). For example, consider in this study the MRI exam with the cutoff between probably positive and equivocal test results and suppose that the MRI test result is going to play the decisive role in selection for surgery. The cross tabulation of test results versus true disease status in the verified sample is given in table 3a together with the number of unverified cases.

Table 3. ILLUSTRATION OF CORRECTION FOR VERIFICATION BIAS. Cross tabulation of test results versus true disease status in a) the study sample and b) the estimated source population.

(a) STUDY SAMPLE

	MRI positive	MRI negative	total
diagnosis positive	11	4	15
diagnosis negative	2	31	33
total verified	13	35	48
unverified	0	10	10
total	13	45	58

(b) ESTIMATED SOURCE

	MRI positive	MRI negative	total
diagnosis positive	11/1.00 = 11	4/0.78 = 5	16
diagnosis negative	2/1.00 = 2	31/0.78 = 40	42
total	13	45	58

Considering the verified cases only we calculate sensitivity, specificity and the predictive values as follows:

sensitivity	=	11/15	=	0.73
specificity	=	31/33	=	0.94
predictive value positive	=	11/13	=	0.85
predictive value negative	=	31/35	=	0.89

From the table it is clear that although 10 negative test results are left unverified all the positive test results are verified. This is intuitively sensible since if the result of the MRI is negative the ophthalmologists will not be keen to operate and any further invasive procedures will be unjustified. The probability that the diagnosis is verified can be expressed conditional on the test result as follows:

probability of verification given a positive test = 13/13 = 1.00
probability of verification given a negative test = 35/45 = 0.78

Knowing that 100% of the positive test results and 78% of the negative test results are verified, we can estimate what the source population (including the unverified cases) might look like. If a sample is a fraction f of the source population, the source population will be 1 / f of the sample. In other words, to estimate the cell frequencies of the source population we divide the cell frequencies of the positive test results by 1.00 and of the negative test results by 0.78. Note that in estimating the source population in this way we assume that the probability of verification depends only on the test result and not on the true disease status. The cross tabulation of the estimated source population is given in table 3b.

From the corrected table we calculate sensitivity, specificity and the predictive values as follows:

sensitivity	=	11/16	=	0.69
specificity	=	40/42	=	0.95
predictive value positive	=	11/13	=	0.85
predictive value negative	=	40/45	=	0.89

Comparing the corrected test characteristics to those calculated from the verified sample we note that the sensitivity is actually lower and the specificity is higher than the verified sample alone suggested. In ROC-curve-terminology the true positive rate (TPR) and false positive rate (FPR = 1-specificity) are both lower than the verified sample suggested. However the predictive values are the same for the verified sample and estimated source population.

The demonstrated correction method can easily be extended to more test results than the dichotomous case, to more diagnostic tests and to include clinical features. With 5 test results the same arithmetic is used but instead of correcting a 2 x 2 table we correct a 2 x 5 table. With k test results and n tests we would correct a 2 x k x n table. The correction was performed firstly assuming the test under consideration plays the decisive role in selection and secondly assuming all 3 tests simultaneously influence selection for verification.

III. RESULTS

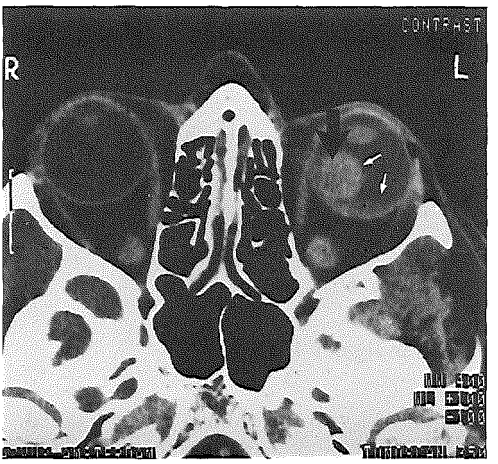
Table 1 summarizes the definitive diagnoses made, the frequency of each diagnosis and, where relevant, the pattern of signal intensities on MRI T1- and T2-weighted images. The patterns of signal intensities on MRI found in this study are in agreement with those reported in previous papers.

Some illustrative examples of orbital SOL are presented in figures 1 through 4.

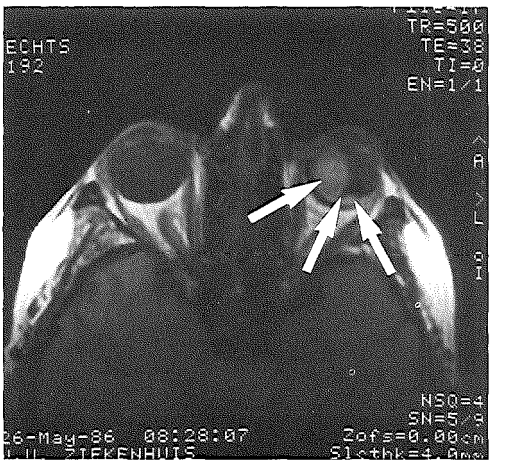
Figure 1 shows CT and MRI images and macroscopy of a 66-year-old woman who presented with a retinal detachment. Clinically a malignant ocular melanoma was suspected (probably positive on the likelihood scale). Both CT and MRI show the typical "butterfly pattern" of a retinal detachment involving the optic disc. On CT the tumor appears as a relatively hypodense area after administration of intravenous contrast. On MRI signal intensities are seen typical for melanotic melanoma, ie. intermediate signal intensity on T1- and low signal intensity on T2-weighted images. The diagnosis "malignant SOL" was considered almost definitively positive on both CT and MRI. The images of both techniques correlate well with the macroscopic findings, which shows a melanoma and retinal detachment.

Figure 2 shows CT, MRI and macroscopy of an infant clinically suspected of having a retinoblastoma although persistent hyperplastic primary vitreous (PHPV) could not be excluded. CT and MRI show a microphthalmus on the left side. On CT a faint hyperdense triangular structure is seen in the middle of the vitreous and PHPV was considered probable. MRI shows an area of intermediate signal on T1- and high signal intensity on T2-weighted images laterally in the globe which made the reading radiologist at the time of the examination report retinoblastoma as being likely. Histology however shows PHPV. The area of intermediate and high signal intensity in the globe on MRI images correlates with haemorrhage as shown by histology. Retrospectively the MRI images would not be considered suspect for retinoblastoma as the lesion does not show the typical signal pattern for retinoblastoma and because of the presence of a microphthalmus. This case constitutes one of the false positive MRI results.

Figure 3 shows images of a 76-year-old woman presenting with an exophthalmus on the right side. Clinically a retrobulbar lesion was suspected, malignancy being equivocal. Both CT and MRI show a retrobulbar lesion. On CT a benign lesion was considered probable because of the smooth contours and homogeneous contrast enhancement. MRI shows a lesion with low signal intensity on all images. However MRI signal intensities do not allow differentiation between malignant and benign optic nerve tumors and malignancy was therefore considered equivocal. Histology proved this to be a benign histiocytoma.

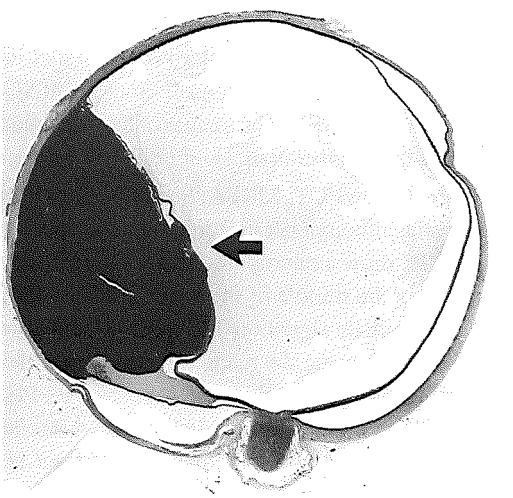


(1a)



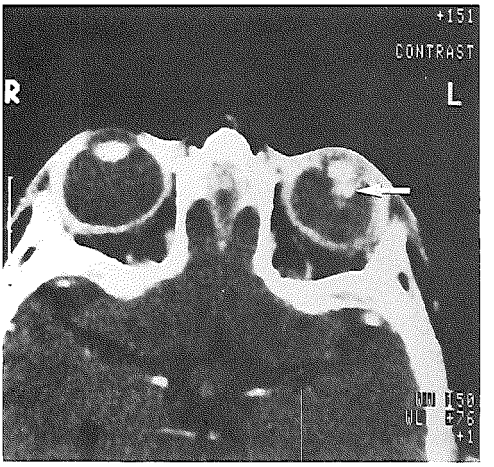
(1b)

(1c)

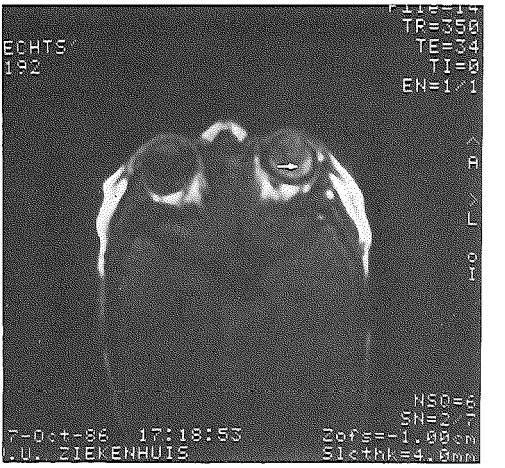


(1d)

Figure 1. Images of a 66-year-old woman who presented with retinal detachment on the left side. Clinically an ocular melanoma was probable. a) The CT image after i.v. contrast administration shows an intraocular lesion ( ) and a retinal detachment ( ). b) The MRI T1-weighted image shows the butterfly pattern of retinal detachment ( ). c) The MRI T2-weighted image shows an area with low signal intensity ( ). d) Macroscopy demonstrates retinal detachment with underlying melanotic melanoma ( ).

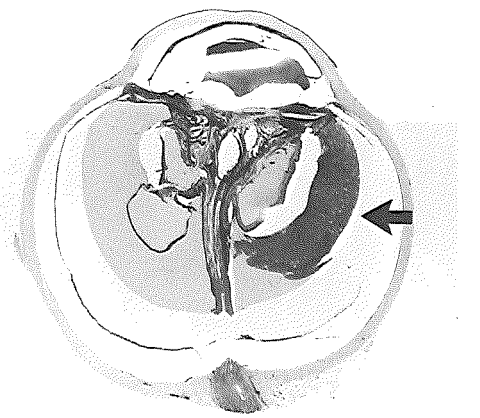


(2a)



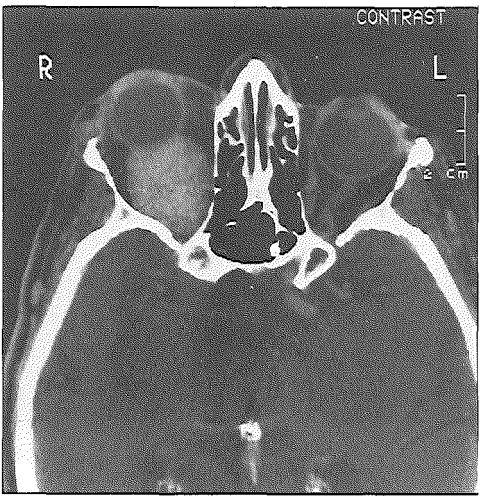
(2b)

(2c)



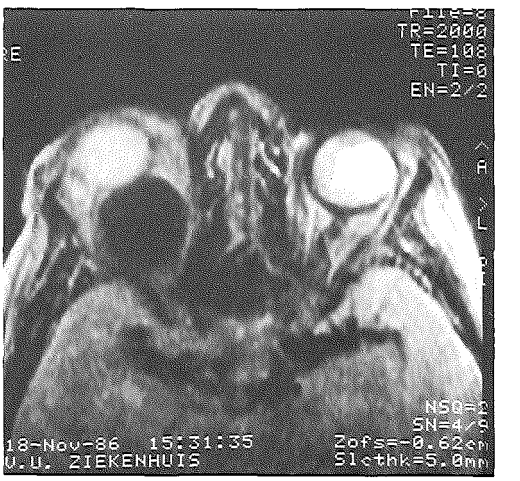
(2d)

Figure 2. 1-month-old infant presenting with a leukokoria on the left side. Clinically retinoblastoma was likely, however persistent hyperplastic primary vitreous (PHPV) could not be excluded. a) The CT image after i.v. contrast administration shows microphthalmus, higher density of the vitreous compared to the normal eye and a faint hyperdense triangular structure in the vitreous (⟵). b) & c) The MRI T1- and T2-weighted images respectively show microphthalmus and an abnormal signal intensity of the vitreous compared to the normal eye. Laterally in the globe an area of intermediate signal on T1- (→) and high signal on T2-weighted images is seen (→). d) Macroscopy and histology show PHPV with haemorrhage laterally (⟵).



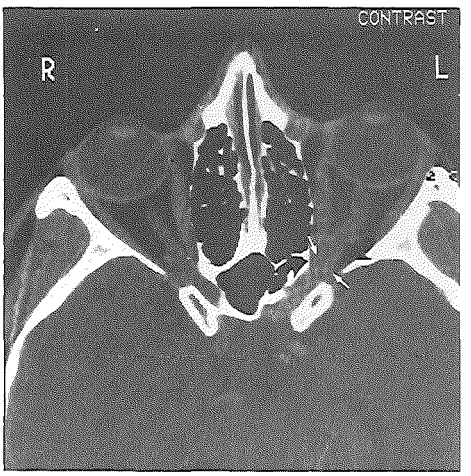
(3a)

(3b)

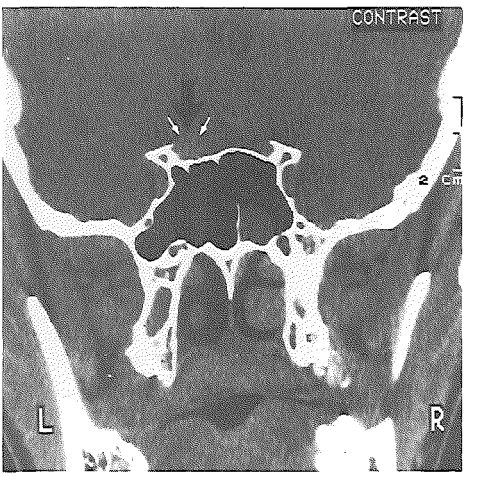


(3c)

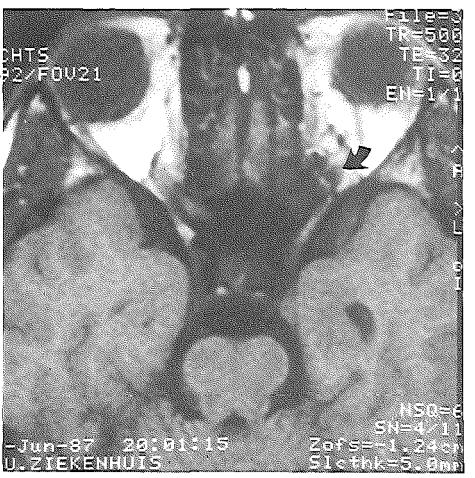
Figure 3. 76-year-old woman presenting with an exophthalmus on the right side. The clinical diagnosis was a retrobulbar lesion, malignancy being equivocal. a) The CT image after i.v. contrast administration shows a homogeneously enhancing retrobulbar lesion. b) & c) The MRI T1- and T2-weighted images respectively show a retrobulbar lesion with low signal intensity on both images. Histology proved this to be a benign histiocytoma.



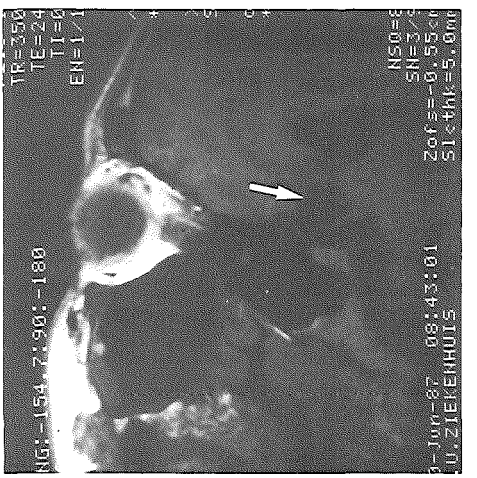
(4a)



(4b)



(4c)



(4d)

Figure 4. 24-year-old man with decreased visual acuity of the left eye. a) The CT axial image after i.v. contrast administration shows a slightly enlarged optic nerve and a small enhancing lesion in the apex of the orbit (↘). b) The CT coronal image after i.v. contrast administration shows a small enhancing area in the region of the anterior clinoid (↘). c) The MRI axial T1-weighted image shows a vascular lesion in the apex of the orbit (◢). d) The MRI oblique T1-weighted image shows a vascular lesion in the region of the anterior clinoid (⟶).

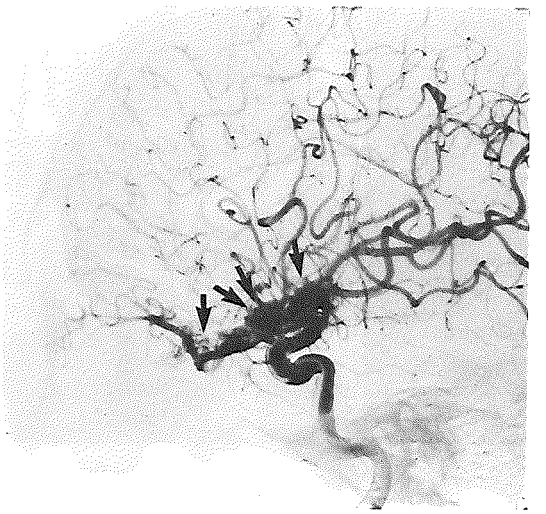


Figure 4. e) Selective left internal carotid angiography (lateral projection) confirmed the presence of an arteriovenous malformation (➡).

Figure 4 shows images of a young man presenting with decreased visual acuity of his left eye. A history of trauma to his left eye was noted. CT shows an enlarged optic nerve (possibly a glioma) and probably a vascular tumor in the apex of the orbit as well as an area of faint contrast enhancement intracranially in the region of the sella. MRI shows a vascular tumor apical in the orbit and in the region of the sella (almost definitively positive on the likelihood scale). The definitive diagnosis was made by angiography which shows an arteriovenous malformation apical in the orbit extending intracranially.

Table 2 gives the cross tabulation of the test results versus the final diagnosis, including the unverified cases. Note that the unverified cases have equivocal, probably negative or negative test results. All cases with positive or probably positive test results are verified.

Figure 5 shows the ROC curves of the clinical, CT and MRI diagnosis without and with correction for verification bias. Table 4 summarizes the areas under the ROC curves for the uncorrected data, the data corrected for verification bias for one test and corrected for all three tests simultaneously. With and without correction for verification bias the ROC curves of MRI and clinical evaluation, and the corresponding areas under the curves, are very similar. The areas under the ROC curves of MRI and clinical evaluation are always larger than that of CT, however no

Table 4. RESULTS. Calculated areas under the ROC curves, the standard errors and p-values in comparing the tests. The p-values tabulated are corrected for correlation between test results.

VERIFIED CASES ONLY

diagnostic test	area under the ROC curve	standard error	comparison between	p-value
clinical	0.95	0.03	clinical-CT	0.18
CT	0.90	0.06	CT-MRI	0.19
MRI	0.95	0.03	clinical-MRI	0.50

CORRECTED FOR VERIFICATION BIAS FOR ONE TEST (ie. assuming selection for verification is determined by the test itself)

diagnostic test	area under the ROC curve	standard error	comparison between	p-value
clinical	0.94	0.03	clinical-CT	0.17
CT	0.88	0.07	CT-MRI	0.15
MRI	0.95	0.03	clinical-MRI	0.39

CORRECTED FOR VERIFICATION BIAS FOR ALL THREE TESTS (ie. assuming all three tests simultaneously influence selection for verification)

diagnostic test	area under the ROC curve	standard error	comparison between	p-value
clinical	0.95	0.03	clinical-CT	0.21
CT	0.90	0.07	CT-MRI	0.23
MRI	0.95	0.03	clinical-MRI	0.50

statistically significant difference can be demonstrated for the uncorrected curves. Likewise, after correcting for verification bias, assuming all 3 tests influence selection for verification, no statistically significant difference can be demonstrated. Assuming that only the test under consideration influences selection for verification the area under the ROC curve of MRI is larger than that of CT at a statistical significance level of p=0.15 using a one-sided test.

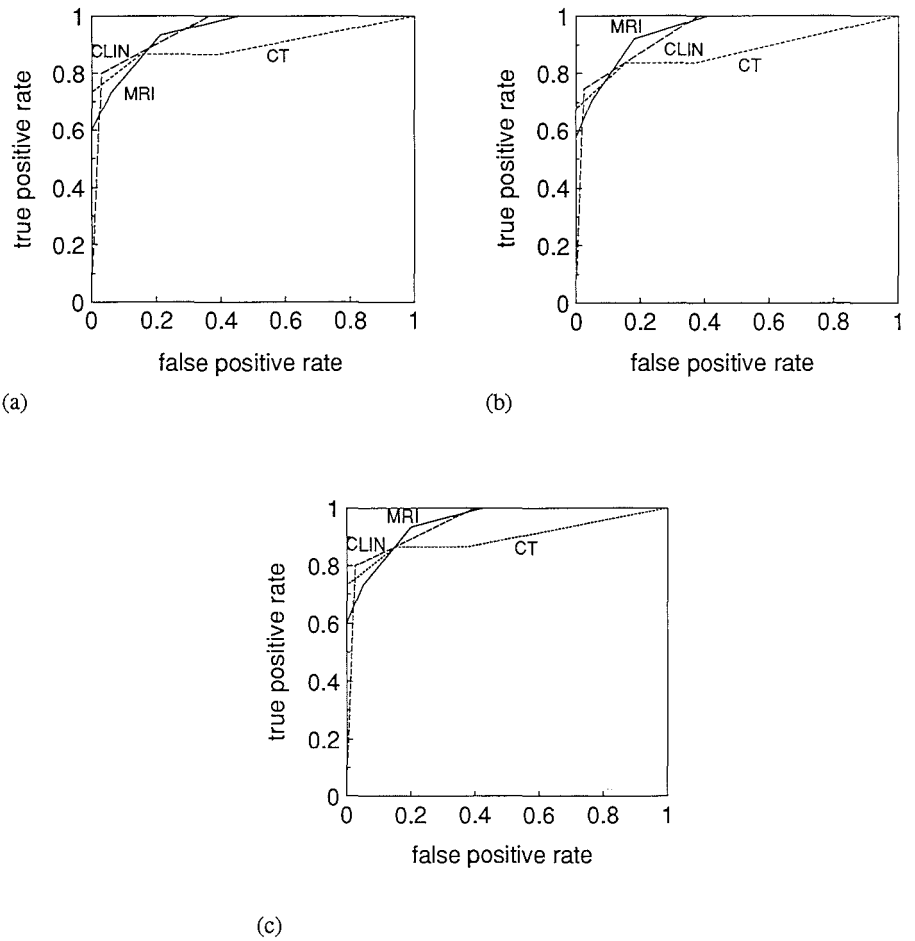


Figure 5. Receiver operating characteristic curves for the clinical diagnosis, CT diagnosis and MRI diagnosis a) uncorrected ROC curves b) corrected for verification bias for one factor (ie. the test under consideration) c) corrected for verification bias for multiple factors (ie. for all three tests)

IV. DISCUSSION

For orbital SOL the clinical, CT and MRI diagnoses were compared using receiver operating characteristic (ROC) methodology. The likelihood of the diagnosis malignant SOL of the orbit or extension of an orbital malignancy was assessed taking into account the clinical presentation. The CT and MRI images were read in the normal everyday routine. The result is therefore a measurement of the performance of the diagnostic system in the clinical context, including the reader and his/her day-to-day variations of performance (10), rather than the diagnostic test alone. The above implies that the conclusions from this study are only valid for the clinical context in which the study was set. The diagnostic procedures were always done in the same sequence, first clinical evaluation then CT then MRI. Our objective was to assess the increment in information after doing CT or MRI as compared to the information provided by clinical evaluation alone.

One of the false positive MRI results was a case of persistent hyperplastic primary vitreous (PHPV) which was diagnosed as retinoblastoma (see figure 2). In retrospect the diagnosis retinoblastoma would not be made on the basis of the MR images and a similar case which presented a few months later was diagnosed correctly. We have not adjusted the results for this "learning effect" assuming that such mistakes tend to occur in the daily routine even after an initial learning phase.

Histology, surgery or angiography were considered the golden standard. However, not all cases could be verified because performing an invasive procedure would have been unethical in the cases concerned. A mathematical correction for this verification bias was applied. The key assumption of this method is that selection for verification is determined by "visible factors" as the test result and clinical presentation. This is a valid assumption because at the time of workup the definitive diagnosis is unknown and the decision to operate or do another procedure is determined by the test results and clinical findings.

To construct the ROC curves we divided the diagnostic truth into two categories. The choice of the division should be clinically relevant, preferably determining the choice of treatment. In this study the diagnostic truths were malignant SOL of the orbit or extension of a malignant orbital lesion versus no malignant disease of the orbit. The diagnosis malignant SOL of the orbit implies a poor prognosis for survival without treatment: enucleation of the orbit should be performed. A benign SOL of the orbit does not always require such an aggressive approach: aggressive treatment is indicated only if vision is threatened. However, there is some overlap between the two chosen groups: adult optic glioma, histologically a benign lesion, has a poor prognosis for both vision and survival and should probably be treated more aggressively than a benign lesion (1).

No statistically significant difference in test performance of either CT or MRI compared to clinical evaluation could be demonstrated suggesting that clinical presentation, taking into account clinical features as age, is a good predictor in evaluating orbital lesions. However, the group studied was quite limited so that a small difference in area under the ROC curves could go undetected (9). Furthermore, two factors probably augmented the performance of clinical evaluation. Firstly, all patients in our series could be examined very well by means of ophthalmoscopy, which is not always the case in the daily routine. Secondly, our series does not include many potential pitfalls for clinical evaluation for example retinal detachment with haemorrhage can clinically easily be confused with melanoma whilst MRI can theoretically distinguish the two.

In this analysis we have concentrated on the incremental value of CT and MRI over clinical evaluation alone and found that by doing a CT or MRI no extra information was obtained. With increasing budget constraints it could be valuable to assess the incremental information provided by expensive technologies as CT and MRI over and above the information provided by clinical evaluation alone instead of only comparing imaging techniques. This may lead the way to more efficient use of modern technology.

In conclusion, although MRI performs better than CT, neither MRI nor CT can be demonstrated to provide significantly more information than clinical evaluation in the diagnosis of orbital SOL. Further assessment of the test performance of CT and MRI could provide us with a limited set of indications for using these technologies in the evaluation of orbital SOL.

V. ACKNOWLEDGMENT

We thank Prof. C.B. Begg, PhD, Prof. Dr. J. Valk and Prof. Dr. C.M.J. Velzeboer for their valuable advice and the technicians of the neuroradiology and MRI departments for their assistance.

VI. REFERENCES

1. Azar-Kia B, Naheedy MH, Elias DA, Mafee MF, Fine M: Optic nerve tumors: Role of magnetic resonance imaging and computed tomography. Radiol Clin North Am 1987; 25.3: 561-581.

2. Bamber D: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psych 1975; 12: 387-415.

3. Begg CB, McNeil BJ. Assessment of Radiologic tests: control of bias and other design considerations. Radiology 1988; 167: 565-569.

4. Bilaniuk LT, Atlas SW, Zimmerman RA: Magnetic resonance imaging of the orbit. Radiol Clin North Am 1987; 25.3: 509-528.

5. Centor RM, Schwartz JS: An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. Med Decis Making 1985; 5: 149-156.

6. Centor RM: A Visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. Med Decis Making 1985; 5: 139-148.

7. Gomori JM, Grossman RI, Shields JA, Augsburger JJ, Joseph PM, DeSimeone D: Choroidal Melanomas: Correlation of NMR spectroscopy and MR imaging. Radiology 1986; 158: 443-445.

8. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839-843.

9. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating charateristic (ROC) curve. Radiology 1982; 143: 29-36.

10. Metz CE: ROC methodology in radiologic imaging. Invest Radiol 1986; 21: 720-733.

11. Begg CB, Greenes RA: Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 207-215.

12. Gray R, Begg CB, Greenes RA: Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Med Decis Making 1984; 4: 151-164.

13. Mafee MF, Putterman A, Valvassori GE, Campos M, Capek V: Orbital space-occupying lesions: Role of computed tomography and magnetic resonance imaging. Radiol Clin North Am 1987; 25.3: 529-559.

14. Sullivan JA, Harms SE: Surface-coil MR imaging of orbital neoplasms. Am J Neuroradiol 1986; 7: 29-34.

Chapter VIII

TESTING FOR FETAL PULMONARY MATURITY: AN ROC ANALYSIS INVOLVING COVARIATES, VERIFICATION BIAS and COMBINATION TESTING[1]

I. INTRODUCTION
II. METHODS
1. The study population and baseline ROC analysis
2. Covariates
3. Verification bias
4. Combinations of tests and prediction rules
5. Technical details
III. RESULTS
1. The baseline ROC curves
2. Covariates
3. Correcting for verification bias
4. Combination testing and prediction rules
IV. DISCUSSION
V. REFERENCES
VI. APPENDIX

ABSTRACT

The L/S ratio and the SPC measurement, amniotic fluid tests performed to assess fetal pulmonary maturity, are evaluated with an ROC analysis. The effect of covariates on the ROC curves is analyzed with a regression methodology, which takes into account all the available data when constructing an ROC curve for each subgroup. To correct for verification bias we use a logistic regression analysis to model the probability of verification, thereby permitting correction for verification bias of a fully stratified data set in spite of small cell frequencies. We examine combination testing with prediction rules using prospective logistic modelling, including as variables test results and clinical features.

---

[1]Co-authors: Douglas K Richardson[1], Peter M Doubilet[2], Colin B Begg[3]. [1]Joint Program in Neonatology and [2]Dept of Radiology, Brigham and Women's Hospital, Harvard Medical School, and [3]Division of Biostatistics and Epidemiology, Dana-Farber Cancer Institute, Boston.

Submitted for publication.

The L/S ratio is a significantly better test than the SPC test. For high gestational age the L/S and SPC tests perform better than for low gestational age. Contamination of the specimen degrades the ROC curves. Correcting for verification bias does not influence the ROC curves significantly but changes the cutoff value of the test variable for any particular operating point. The use of prediction rules to evaluate combination testing shows that performing the SPC test in addition to the L/S test adds no significant information compared to performing only the L/S test. Including gestational age in the prediction rule of either test improves the prediction.

I. INTRODUCTION

An infant with immature lungs has a risk of developing respiratory distress syndrome (RDS) of the newborn. Faced with the decision to continue or prematurely deliver a complicated pregnancy, obstetricians often use amniotic fluid testing to assess pulmonary maturity, thereby permitting delivery of infants at low risk for respiratory distress syndrome (RDS). The lecithin/sphingomyelin (L/S) ratio is the biochemical test most used. Numerous alternative amniotic fluid tests, such as the saturated phosphatidyl choline (SPC) assay, have been reported but none have come into widespread use.

Published studies concerning the L/S ratio and related tests (7,19,25) report sensitivity and specificity using a single conventional cutoff value to distinguish between diseased and non diseased. Unfortunately, the use of numerous variations in assay technique for the L/S ratio by different investigators makes direct comparisons between studies meaningless (7). Furthermore, the appropriate cutoff value in practice depends on the clinical context and patient population in which the test is used. Receiver operating characteristic (ROC) methodology circumvents these problems, as it presents pairs of true positives and false positives for the full range of possible cutoff values, and the ROC indices do not depend on the choice of a particular operating point (16). The area under the ROC curve is a convenient summary measure of test performance and is useful for comparing tests (10,11,16).

The use of ROC analysis, however, is not a panacea. A number of methodological problems exist which are the focus of this paper. Covariates can have a major impact on the ROC curve (20). Various forms of bias can influence the evaluation of test performance, of which verification bias is an important one (3). Furthermore, when multiple diagnostic tests are available it is of interest not only to choose the best test but also to consider combination testing.

In testing for fetal pulmonary maturity, gestational age of the fetus is a critical covariate (26). Investigators reporting the evaluation of the L/S ratio and related tests have generally assumed that sensitivity and specificity do not change with increasing gestational age (7,25,26). In view of the biological changes of lung maturation with increasing gestational age, this assumption is probably incorrect (23). However, in stratifying the data set for the covariate gestational age,

the subgroups are likely to become so small that calculations become meaningless. To cope with this problem, we use a general regression methodology (24) to determine the effect of gestational age on the ROC curves of the L/S ratio and related tests.

Contamination of the amniotic fluid specimen with blood or meconium can be considered a second covariate. Because contaminated specimens give inaccurate results, these have conventionally been eliminated from the analysis when evaluating fetal lung maturity (25). However, as many as 20 % of specimens are contaminated. Obtaining no results from such a substantial proportion of tests is a serious clinical handicap, so a few centers assay the specimens and report the results accompanied by cautionary disclaimers. Therefore, the clinician needs a measure of how good the test is for uncontaminated and contaminated specimens separately. Instead of disregarding the contaminated specimens, we constructed a separate ROC curve for these specimens.

Many types of bias can influence the estimates of the test parameters, such as verification bias, uninterpretability bias, test review bias and absence of a definitive gold standard (3,20). Of these biases, verification bias is of major concern in this analysis.

Verification bias exists when not all subjects tested undergo the definitive procedure to assess true disease status (1). In testing for fetal pulmonary maturity the gold standard is the development of respiratory distress syndrome (RDS) of the newborn and delivery is therefore the definitive procedure to assess true disease status. The result of amniotic fluid testing is considered verified if the baby is delivered within a set time limit after collection of the specimen, conventionally 48 to 72 hours. Waiting longer would mean that the lungs can mature further and the test result would no longer be representative of pulmonary status at birth. If the test results indicate immaturity of the fetus, every effort will be undertaken to delay delivery, thus resulting in a test result which has not been verified by the gold standard. Delivery of a fetus with an immature test result usually occurs from circumstances beyond the clinician's control. Immature (positive) test results are thus less likely to be verified than mature (negative) test results. Consequently, true and false positive test results will be missed to a greater extent than true and false negatives, implying that the true positive and false positive rates calculated from the verified sample will be biased in that they are underestimates of the actual rates. That is, the sensitivity (true positive rate) will be underestimated and the specificity (1-false positive rate) will be overestimated. A mathematical correction (1) for this bias can be applied to the data set.

Faced with different diagnostic procedures to test for a disease, it is important not only to choose the best test but also consider combinations of tests. Although combination testing of amniotic fluid tests has been studied (9,14), the problem has not been addressed using ROC methodology. Various approaches are possible to select an optimal combination of tests taking clinical features into account (5,12). A convenient approach is to create prediction rules based on clinical features and test values by means of regression analysis. Creating prediction rules is not novel in fetal pulmonary maturity testing (18,23), but ROC curves of the prediction rules have not been constructed and gestational age has only sporadically been included as a variable (23). In studying

fetal pulmonary maturity tests we developed a prediction rule based on the gestational age, the L/S ratio and SPC test. With multivariate regression techniques and ROC curves of the prediction rules we determined which tests add significant information to the predicted probability of RDS.

II. METHODS

1. The study population and baseline ROC analysis

The log book of the Fetal Lung Maturity Laboratory of the Boston Lying-In Hospital and subsequently the Brigham and Women's Hospital was used to identify all women who had delivered within 72 hours following amniotic fluid testing in the years 1979 to 1983 and calender year 1987. Specimens were analyzed for L/S according to the technique of Gluck and for SPC according to the technique of Torday, and were processed regardless of contamination with blood or meconium. Mother's and infant's charts were abstracted. All newborns requiring oxygen were reviewed by a single reviewer, as were all deaths. In addition, test results of fetuses tested but not delivered within 72 hours were recorded during the period 1981 to 1982. In total 854 verified cases and 653 unverified cases were recorded.

A positive test in this setting is a test result that suggests pulmonary immaturity (ie. that predicts the development of respiratory distress syndrome of the newborn.) The "gold standard" is the development of RDS in the newborn.

The definitive diagnosis or outcome is one of two disease classes: disease (respiratory distress syndrome of the newborn (RDS)); no disease (no RDS). Respiratory problems such as transient tachypnea of the newborn (TTN), ventilatory support for apnea, chronic pulmonary insufficiency of extreme prematurity, pneumonia and neonatal hydrops were counted as "no RDS". Infants with congenital anomalies of the cardiopulmonary system were excluded. Eleven patients had an atypical clinical picture of RDS and these were omitted from the baseline ROC analysis. Also omitted from the baseline analysis were uninterpretable test results, contaminated specimens and unverified cases.

2. Covariates

Covariates may affect the performance, and thus the ROC curves, of diagnostic tests (20,24). The simplest approach to analyze this effect is to stratify the data set into subgroups. In this analysis gestational age is the major covariate. We stratified the data set into subgroups by gestational age as follows: ≤ 28 weeks; > 28 but ≤ 32 weeks; > 32 but ≤ 36 weeks; > 36 weeks.

Contamination of the amniotic fluid specimen with blood or meconium is a second covariate. If the specimen is contaminated by blood or meconium the test is not entirely uninterpretable, but the test result may be altered by the contaminant. In particular, blood and meconium have an L/S ratio of about 2, and so will change the specimen's L/S ratio towards this value. The L/S ratio of contaminated specimens may, therefore, be less adequate in discriminating disease from non-disease. Theoretically, contamination should not influence the SPC measurements. We subdivided the data into contaminated specimens (either blood or meconium) and uncontaminated specimens and constructed ROC curves for both type of specimens.

A problem of dividing the data by covariates is that the subgroups are often so small that the ROC curves and related statistics become too imprecise to interpret reliably. Furthermore, if the covariate is continuous, this property will not be accounted for. The use of a regression methodology is helpful in adjusting for covariates (15,24). Such an approach allows one to additionally adjust for other covariates. We implemented the regression methodology using the software package PLUM (15,24). PLUM is an interactive program for the analysis of ordinal data, which may be used to model ROC curves. The data is entered as a frequency table, together with the covariates, and interactively a model is fitted. The program produces parameter estimates that define the model, which are then used to generate the ROC curve.

3. Verification bias

Workup bias or verification bias (1,8,20) exists when not all subjects tested undergo the definitive procedure to assess true disease status. The test result, often in combination with clinical features, can influence the decision to verify the diagnosis.

An amniotic fluid test is considered verified if delivery of the baby occurs within 72 hours of specimen collection. If the baby is delivered later than 72 hours after collection of the specimen further maturation will occur and the test is not representative: such a test result is therefore unverified, meaning that there is no gold standard against which to determine the correctness of the test result for such a case. In the setting of this analysis the selection for verification (delivery) is greater among negative (mature) test results than among positive (immature) test results (3), and the true positive and false positive rates calculated from the verified sample will be underestimates of the actual rates. The decision to delay delivery (if at all possible) will be influenced by gestational age and the test results.

The mathematics of the correction for verification bias is best explained with an example, which is given in Appendix 1. The basic assumption is that selection for verification (and thus the probability of verification) depends on clinical factors and test results, and does not depend directly on true disease status. An equivalent assumption is that the predictive values are the same for the verified and source population (1,8). As shown in appendix 1, if we subdivide the

population into cells containing the frequency of a particular combination of clinical features and test results, the estimated frequency of that cell in the source population is the cell frequency in the verified sample divided by the probability of verification for that cell.

Assuming that gestational age, the L/S test and the SPC test together influence the decision to deliver the baby, we need to correct the data set for all three. The estimated frequency in the projected source population of a particular combination of gestational age, L/S and SPC results, is the frequency in the verified sample, divided by the probability of verification given that combination.

An arithmetical problem arises when we attempt to determine the probability of verification by subgrouping into categories for gestational age and the two test results: the subgroups are very small and some cells are empty. A solution to this problem is to perform a logistic regression analysis on the probability of verification, which takes into account all the available data. We have chosen to do the regression analysis using the values of the variables on the continuous scale (not categorized). Subsequently we corrected the frequency tables for each test by dividing by the expected probability of verification as determined by the regression model. This corrects the data of each test for verification bias caused by multiple factors: gestational age, L/S ratio and SPC measurement.

The standard error of the area under the corrected ROC curve can never be smaller than that of the baseline ROC and the greater the proportion of unverified cases present, the more uncertain we will be of the actual state of affairs. The corrections for verification bias inflate the apparent sample size and will result in an underestimate of the standard error. Furthermore, because the unverified cases were recorded during only one third of the total study, we counted each unverified case as approximately 3 cases. This further inflates the apparent sample size. To estimate the standard error of the corrected curve we deflated our projected source population to be equivalent in total size to the verified sample. To validate this crude estimate we also calculated the standard error using formulas published elsewhere (1).

4. Combinations of tests and prediction rules

We used logistic modelling to derive a prediction rule for RDS. The dependent, or predicted, variable was the probability of RDS, and the independent variables were gestational age, the L/S ratio and the SPC measurement. Subsequently, prediction rules were derived using as variables 1) gestational age only, 2) the L/S ratio only, 3) the SPC measurement only, 4) gestational age and the L/S ratio, 5) gestational age and the SPC measurement and 6) the L/S ratio and the SPC measurement. Only verified cases were included to estimate the parameters of the prediction rules. With standard regression techniques we determined if a variable was significant in the regression equation compared to the (less complex) model without it (using the Likelihood ratio test (13,22)). The predicted probability of disease was then used as a combination test result and an ROC curve of the prediction rule was constructed by using different cutoff points on the

predicted probability scale. For example, for a cutoff point p the test is positive if the predicted probability is $\geq p$ and negative if the probability $< p$. The difference between the ROC curve based on the prediction rule of gestational age and both tests compared to the ROC based on gestational age and only one test, is a measure of how much information is obtained from performing the extra test.

5. Technical details

The test results in this analysis are on a continuous rather than an ordinal scale. We have chosen to treat the data as ordinal where applicable. The test results were divided into 15 categories. For the analyses of subgroups we divided the test results into 10 categories, since the sample size is smaller in these cases. A more refined division will not give much more information (21). The cutoff points were chosen in such a way that the pairs of true positive and false positive rates are fairly evenly distributed along the ROC curve.

The Mann Whitney U statistic (equivalent to the trapezoidal rule) was used to calculate the area under the ROC curve (4,10). We checked a number of our calculations with the Dorfman Alf algorithm (6) and found that with 10 to 15 cutoff points that are evenly distributed the two methods give practically the same results. The standard error of the area was calculated using the Hanley-McNeil algorithm (10). Smooth curves were constructed with a least squares estimate of the normal deviate values of the true and false positive rates.

We used a two-sided paired t-test to compare the areas under two ROC curves, taking into account the correlation between test results performed on the same cases (11). The correlation coefficient used in the t-test was found in table I of reference 11, which is a coefficient representing the correlation between the areas under two ROC curves derived from the same cases.

The ROC indices of the baseline curves and the curves stratified for covariates, as well as the corresponding t-test statistics for comparing the curves, were calculated using standard techniques as described above. After applying the regression methodology for gestational age, we calculated the indices and statistics using the standard techniques, assuming that the correlation between test results derived from the same cases is the same as without modelling. The significance of gestational age on the ROC curves was calculated from the parameter estimates of the regression model. The data corrected for verification bias was treated as if it were raw data, after having deflated the projected source population to be equivalent in total size to the verified sample (as described in paragraph 3). For the comparative statistic between the L/S and SPC test after correction for verification, we used the same correlation coefficient as for the baseline curves. To compare the baseline curve with the curve corrected for verification bias, we used two alternative assumptions: a correlation coefficient equal to 0, and equal to the fraction of verified cases. The significance of including an extra variable in the prediction rule was derived directly from the logistic model.

Table 1. Frequency tables of test results in the verified sample. For illustrative purposes the results have been divided into 5 categories for the L/S and 6 for the SPC, instead of 15 as used in the analysis.

a) L/S ratio

L/S ratio: range	0.1-0.9	1.0-1.7	1.8-2.2	2.3-2.9	3.0-9.9
disease	7	26	7	5	2
no disease	8	53	80	120	363

b) SPC

SPC: range	1 -250	251 -500	501 -750	751 -1000	1001 -1250	>1250
disease	14	18	6	6	1	2
no disease	21	57	104	124	79	239

Table 2. Areas under the ROC curves and their standard errors.

ROC description	L/S area(error)	SPC area(error)
1. Baseline	.90 (.02)	.85 (.03)[1]
2. Effect of covariates		
a) gestational age:		
stratification:		
25 - 28 weeks	.76 (.10)	.65 (.11)
28 - 32 weeks	.84 (.05)	.84 (.04)
32 - 36 weeks	.87 (.04)	.83 (.05)
regression(PLUM)[3]:		
25 - 28 weeks	.70 (.09)	.65 (.10)
28 - 32 weeks	.85 (.04)	.85 (.04)
32 - 36 weeks	.87 (.04)	.80 (.05)
b) contaminated specimens	.84 (.04)	.79 (.07)
3. Corrected for verification bias	.88 (.02)	.83 (.03)[2]

[1] comparing L/S and SPC: p=0.02

[2] comparing L/S and SPC: p=0.005

[3] effect of gestational age: L/S: p=0.0005 SPC: p=0.025

III. RESULTS

1. The baseline ROC curves

Table 1 enumerates the frequencies of the test results versus disease and figure 1 presents the baseline ROC curves for L/S and SPC. These curves are based on all gestational ages, and restricted to uncontaminated specimens and verified cases. The calculated areas (table 2) are 0.90 for the L/S and 0.85 for the SPC test. All subsequent analyses and corrections were done using the baseline ROC's as starting point.

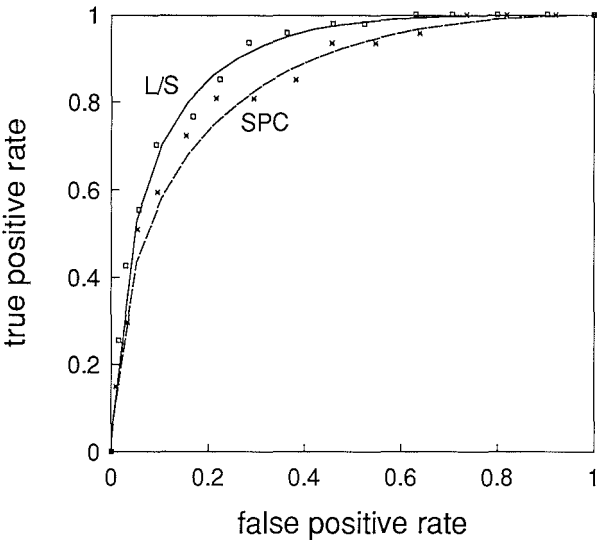
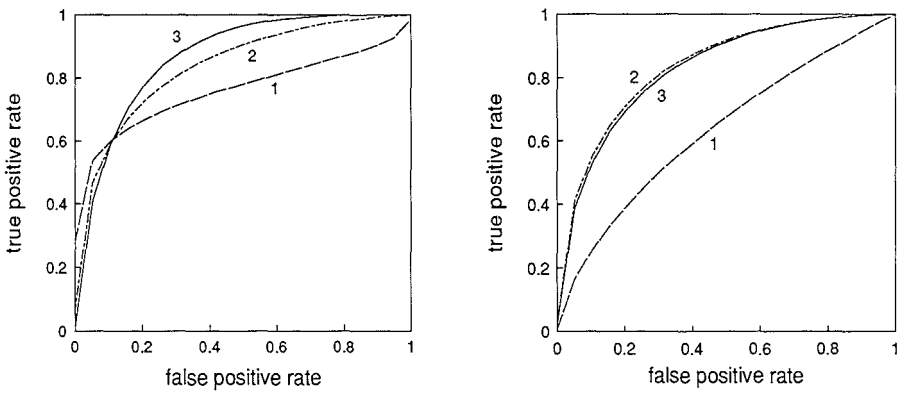


Figure 1. Baseline ROC curves of the L/S ratio and SPC measurement in testing for fetal pulmonary maturity. Observed points (symbols) and smoothed curves.

2. Covariates

a) Gestational age

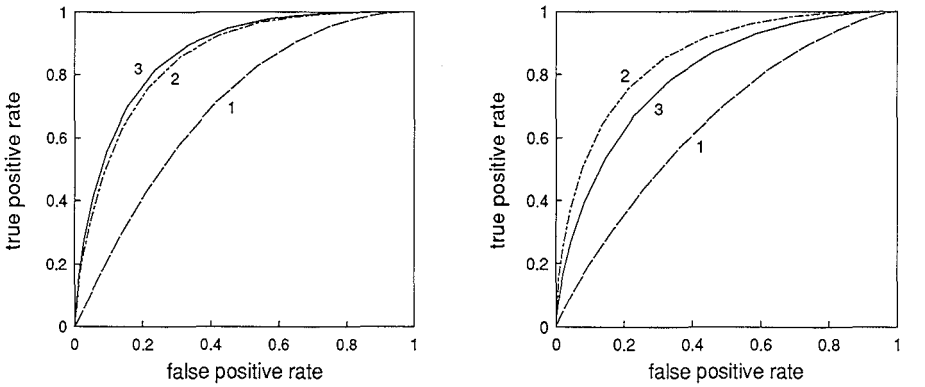
We stratified the data set into subgroups by gestational age. The ROC curves of the subgroups are shown in figure 2 and the indices tabulated in table 2. In the subgroup of gestational ages beyond 36 weeks there are no cases of RDS and it is therefore impossible to draw an ROC curve for this subgroup. For both the L/S and the SPC the area under each of the 3 ROC curves is smaller than the area under the baseline curve because the cases contributing strongly to specificity of the baseline curve are those beyond 36 weeks gestational age.



a) L/S b) SPC

Figure 2. ROC curves for different gestational ages using stratified data set. 1: ≤ 28 weeks, 2: > 28 ≤ 32 weeks, 3: > 32 ≤ 36 weeks.

As can be seen from figure 2 the L/S curve for the youngest gestational age group is not a "proper" ROC curve, that is, the curve is not uniformly convex towards the upper left corner (17). This often happens when drawing curves for small samples. Using the software package PLUM we performed a regression analysis for the covariate gestational age. The resulting curves are shown in figure 3 and the indices are summarized in table 2. The effect of gestational age on the area under the ROC curve is statistically significant (p=0.025), after modelling with PLUM, for both the L/S and the SPC: both improve with increasing gestational age.



a) L/S **b) SPC**

Figure 3. ROC curves for different gestational ages generated by PLUM rgeression technique. 1: ≤ 28 weeks, 2: > 28 ≤ 32 weeks, 3: > 32 ≤ 36 weeks.

b) Contamination

19 % of all verified specimens were contaminated with blood and 3 % were contaminated with meconium. The contaminated specimens were omitted in the baseline analysis, as is customary in reports on fetal pulmonary maturity testing. Although not statistically significant, the ROC curve for each test is worse for contaminated specimens than for uncontaminated specimens (table 2). The results of contaminated specimens, however, provide discriminatory power compared to the null result if the specimen is discarded.

3. Correcting for verification bias

We performed a logistic regression analysis on the probability of verification, modelling the variables gestational age (GA), the L/S ratio (LS) and the SPC value (SPC) using the values on the continuous scale. The logistic regression equation is of the form

Table 3. Frequency tables of test results in the projected source population after correction for verification bias for a) the L/S test and b) the SPC test (compare with table 1).

a) L/S ratio

L/S ratio: range	0.1-0.9	1.0-1.7	1.8-2.2	2.3-2.9	3.0-9.9
disease	42	149	33	21	6
no disease	44	266	341	426	922

b) SPC

SPC: range	1 -250	251 -500	501 -750	751 -1000	1001 -1250	>1250
disease	85	101	26	26	3	9
no disease	114	266	408	406	224	581

$$P(V^+ \mid GA, LS, SPC) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot GA + \beta_2 \cdot LS + \beta_3 \cdot SPC)}}$$

The parameter estimates from the logistic regression analysis are:

$\beta_0 = -4.225$

$\beta_1 = 0.074$

$\beta_2 = 0.179$

$\beta_3 = 0.00030$

(Note that the β's are not comparable because they are in different units.)

To illustrate this regression model, consider a fetus of 28 weeks gestational age with an L/S ratio of 1.8 and an SPC value of 600. Substituting these values in the above equation we determine that a fetus with this combination of test results and clinical features will have a 16% probability of being delivered (in other words the probability of verification is 16%). If the gestational age is 34 weeks, the L/S 6.4 and the SPC 1125 the probability of delivery is 44%. The model predicts that the more mature the test results are, the more likely the tests will be verified by delivering the baby.

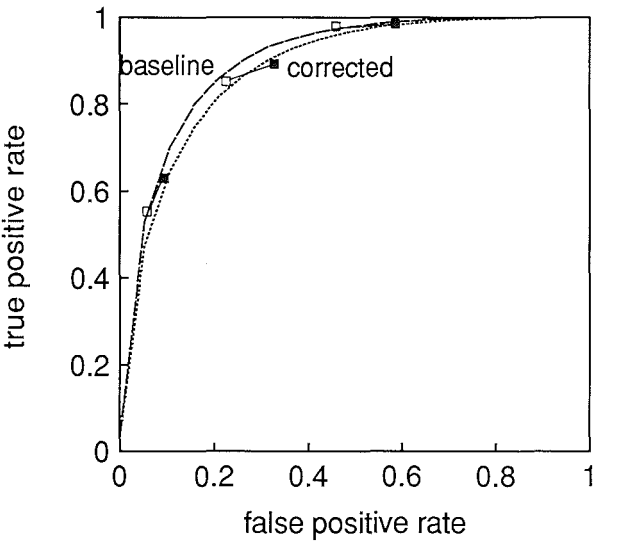
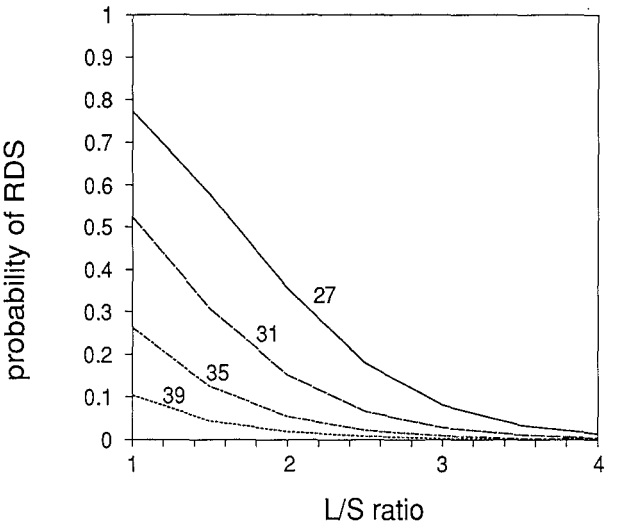
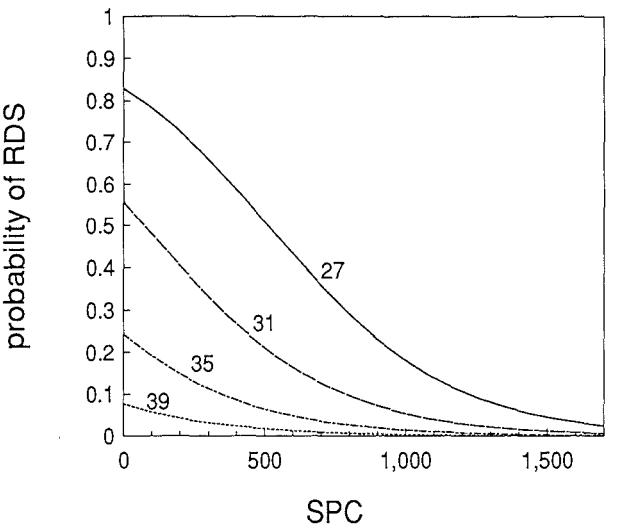


Figure 4. Baseline ROC curve and ROC curve after correction for verification bias for the L/S ratio. Three operating points are indicated to illustrate their shift after correction for verification. Note that these are observed points, that is before the curve is smoothed. Open squares are uncorrected and solid squares are corrected operating points.

We corrected the cell frequencies of the cross tabulation by dividing by the probability of verification as determined by the logistic model, substituting the median values of gestational age, L/S ratio and SPC value for each cell in the above equation. Table 3 gives the frequency table of test results in the projected source population. After correction for verification bias the ROC curves (Figure 4) and the calculated areas (Table 2) are similar to the baseline ROC curves and areas. However, the location of the operating points have shifted up on the ROC curve (Figure 4). This implies that although the shape of the ROC curve is not greatly affected by verification bias in this analysis, the sensitivity and specificity at any particular cutoff value of the test variable are clearly affected (see also Appendix 1).



a) GA and L/S



b) GA and SPC

Figure 5. Predicting RDS based on gestational age and either the L/S ratio or the SPC measurement. The curves shown are for 27, 31, 35 and 39 weeks gestational age.

Table 4. Results of statistical tests of the prediction rules, comparing a less complex model 1 with a more complex model 2. Area under the ROC curves and the standard error of the area. The variables are GA=gestational age, L/S=L/S ratio and SPC=SPC measurement.

variables in model 1	variables in model 2	p value
GA	GA and L/S	<0.0001
GA	GA and SPC	<0.0001
L/S	GA and L/S	<0.0001
L/S	L/S and SPC	0.6
SPC	GA and SPC	<0.0001
SPC	L/S and SPC	<0.0001
GA and L/S	GA, L/S and SPC	0.4
GA and SPC	GA, L/S and SPC	<0.0001
L/S and SPC	GA, L/S and SPC	<0.0001

ROC curves of the prediction rules:	area (error)
GA	.85 (.02)
L/S	.90 (.02)
SPC	.85 (.03)
GA and L/S	.92 (.02)
GA and SPC	.90 (.02)
L/S and SPC	.90 (.02)
GA, L/S and SPC	.92 (.02)

4. Combination testing and prediction rules

The logistic regression equation used to predict RDS on the basis of gestational age (GA), the L/S ratio (LS) and SPC measurement (SPC) is of the form:

$$P(D^+ \mid GA, LS, SPC) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot GA + \beta_2 \cdot LS + \beta_3 \cdot SPC)}}$$

In addition to this comprehensive prediction rule, we derived 6 other prediction rules using as variables:

1) gestational age only
2) the L/S ratio only
3) the SPC measurement only

4) gestational age and the L/S ratio
5) gestational age and the SPC
6) the L/S ratio and the SPC measurement

The logistic regression was performed including the verified cases only. Plotting the predicted probability of RDS as function of the selected variable(s) for the prediction rules 4 and 5, we derive the graphs given in figure 5. The graphs show, as one would expect, that for low gestational age or low L/S or low SPC, the probability of disease is high, and that the predicted probability of disease decreases with an increase in any of the three variables.

Results of the statistical tests and the indices of the ROC curves of the prediction rules are summarized in table 4. The analysis demonstrates that including gestational age as variable into the prediction rule is significant. The L/S test adds significant information even if the SPC test and gestational age are in the prediction rule. However, the addition of the SPC test if the L/S is in the prediction rule is not significant.

IV. DISCUSSION

ROC analysis is a convenient method for comparing diagnostic tests and may provide insight into the practical use of diagnostic tests. However, as mentioned before, it is not a panacea and a number of methodological issues exist. Covariates may influence the ROC analysis considerably. Verification of only a subgroup of the tested subjects may bias the test parameters and ROC curve. Furthermore, when multiple diagnostic tests are available one has to consider combination testing.

ROC analysis of the data on fetal pulmonary maturity testing provides interesting insights into the clinical use of the L/S ratio and SPC measurement. A baseline analysis of all gestational ages, restricted to verified cases and uncontaminated specimens, showed that the L/S ratio is significantly better than the SPC test. As illustrated by the analysis, evaluating diagnostic tests with ROC analysis brings with it a number of methodological issues that can substantially affect the ROC curve and its indices.

The effect of covariates on an ROC curve can be examined in two ways: by dividing the population into subgroups or by use of a regression methodology. The former approach is often impractical, as the sample sizes of the subgroups tend to be too small. Using a regression methodology takes into account all the available data when constructing an ROC curve for each subgroup. This leads to more meaningful ROC curves with smaller standard errors of the calculated ROC indices.

Analyzing the data set stratified for covariates showed that gestational age influences the ROC curves. Both the L/S and SPC tests perform better for higher gestational ages. Analysis of the data set with prediction rules also showed that gestational age should be taken into account when interpreting the test results.

As expected, the test performance of the L/S test for contaminated specimens is worse than the same test with uncontaminated specimens. Contrary to what we expected, the SPC test was also worse for contaminated specimens, although the difference is not significant. The difference in test performance for contaminated specimens between the L/S ratio and the SPC is not statistically significant.

When correcting for verification bias in this multiply stratified data set, there was a problem of small cell frequencies. A logistic regression analysis was used to model the probability of verification, in order to make appropriate corrections.

If verification is subject to selection bias, as is often the case, correcting for this bias may result in a similar or identical ROC curve with the same overall test performance. However, verification bias will change the sensitivity and specificity of any particular cutoff value of the test variable and shift the operating points along the ROC curve. We observed this phenomenon in our analysis. Consider, for example, the routinely used cutoff value of 2 for the L/S test. A cutoff value of 2 for the L/S test corresponds to a point on the baseline ROC curve with slope 0.9. The same operating point on the corrected ROC curve (that is the point with the same slope of 0.9) corresponds to a cutoff value of 1.8 for the L/S test. Correcting this data set for verification bias implies adjusting the chosen cutoff value of the test variable to a stricter criterion. Thus, if we were to perform a utility analysis, the optimal operating point could be affected by verification bias.

A number of other biases are relevant in ROC analysis (2,3). We analyzed the data set to account for uninterpretable test results, test review bias and bias due to absence of an unequivocal gold standard. These three types of bias had little effect on the results and are therefore not presented in detail.

Prediction rules based on test results and clinical features are convenient in analyzing combination testing. Using standard regression techniques one can determine whether the inclusion of a test variable in the prediction rule is significant. ROC analysis of the prediction rules is another method of estimating the additional information obtained by performing an extra test.

The use of prediction rules to evaluate combination testing shows that performing the SPC test in addition to the L/S test adds no significant information compared to the L/S test only. Taking gestational age into account in the prediction rule improves the prediction as compared to using only the test result(s). From the prediction model we derived graphs convenient for clinical use in predicting fetal pulmonary immaturity based on gestational age and either the L/S test or the SPC test (Figure 5). Using the prediction rule of the L/S ratio only, we calculate that the routinely used L/S ratio of 2 corresponds to a predicted probability of RDS of 11%. If we assume that the routinely used cutoff value of 2 for the L/S test corresponds with the desired cutoff value of predicted probability of RDS, then the same predicted probability should apply for different gestational ages. A predicted probability of 11% would correspond, for gestational ages of 27,

31, 35 and 39 weeks, to L/S cutoff values of 2.8, 2.2, 1.6 and 1.0 respectively (Figure 5a). In other words, the higher the gestational age, the stricter the criterion for a positive (immature) L/S ratio should be.

In conclusion, various issues should be considered when evaluating diagnostic test performance with ROC methodology, such as the effects of covariates on test performance, corrections for verification bias and combination testing. In performing an ROC analysis of the L/S ratio and the SPC measurement for fetal pulmonary maturity testing we found that these methodological issues affect results substantially enough to impact clinical decision making. The L/S ratio is a better test than the SPC test. For high gestational age the L/S and SPC tests perform better than for low gestational age. Contamination of the specimen degrades the ROC curves but still provides clinically important information compared to discarding the specimen. Correcting for verification bias does not influence the ROC curves significantly but changes the cutoff value of the test variable. The use of prediction rules to evaluate combination testing shows that performing both the L/S and SPC test adds no significant information compared to performing only the L/S test, and that including gestational age into the prediction rule of either test improves the prediction.

V. REFERENCES

1. Begg CB, Greenes A. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 207-215.

2. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. J Chron Dis 1986; 39-8: 575-584.

3. Begg CB, McNeil BJ. Assessment of Radiologic tests: control of bias and other design considerations. Radiology 1988; 167: 565-569.

4. Centor RM, Schwartz JS. An evaluation of methods for estimating the area under the receiver operating characteristic curve. Med Decis Making 1985; 5: 149-156.

5. Diehr P, Highley R, Dehkordi F, Wood R, Krueger L, Teitz C, Hermanson B. Prediction of fracture in patients with acute musculoskeletal ankle trauma. Med Decis Making 1988; 8: 40-47.

6. Dorfman DD, Alf E: Maximum Likelihood Estimation of parameters of signal detection theory - a direct solution. Psychometrika 1968; 33/1 : 117-124.

7. Freer DE, Statland BE. Measurement of Amniotic Fluid Surfactant. Clin Chem 1981; 27/10, 1629-1641.

8. Gray R, Begg CB, Greenes RA. Construction of Receiver Operating Characteristic Curves when Disease Verification is subject to selection bias. Med Decis Making 1984; 4: 151-164.

9. Hallman M, Arjomaa P, Mizumoto M, Akino T. Surfactant proteins in the diagnosis of fetal lung maturity. I. Predicitive accuracy of the 35 kD protein, the lecithin/sphingomyelin ratio, and phosphatidylglycerol. Am J Obstet Gynecol 1988; 158: 531-535.

10. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic curve. Radiology 1982; 143: 29-36.

11. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148; 839-843.

12. Hu S, Haimes YY, Galen RS. Optimal selection of a battery of tests: a multiobjective optimization methodology. Med Decis Making 1988; 8: 19-32.

13. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and quantitative methods. Lifetime Learning Publications, Wadsworth, Inc. Belmont, California; 1982.

14. Kulovich MV, Gluck L. The lung profile. II. Complicated pregnancy. Am J Obstet Gynecol 1979; 135: 64-70.

15. McCullagh P. Regression models for ordinal data. J R Statist Soc B 1980; 42: 109-142.

16. Metz CE. ROC methodology in radiological imaging. Invest Radiol 1986; 21 : 720-733.

17. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24: 234-245.

18. Nelson GH. Risk of respiratory distress syndrome as determined by amniotic fluid lecithin concentration. Am J Obstet Gynecol 1975; 121: 753-755.

19. O'Brien WF, Cefalo RC. Clinical applicability of amniotic fluid tests for fetal pulmonic maturity. Am J Obstet Gynecol 1980; 136: 135-144.

20. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299: 926-930.

21. Rifkin RD. Maximum Shannon Information content of diagnostic medical testing. Med Decis Making 1985; 5: 179-190.

22. SAS User's guide: Statistics. Version 5. SAS Institute Inc. Cary, North Carolina.

23. Schlueter MA, Phibbs RH, Creasy RK, Clements JA, Tooley WH. Antenatal prediction of graduated risk of hyaline membrane disease by amniotic fluid foam tes for surfactant. Am J Obstet Gynecol 1979; 134: 761-766.

24. Tosteson ANA, Begg CB. A General Regression methodology for ROC curve estimation. Med Decis Making 1988; 8 : 204-215.

25. Tsai MY, Shultz EK, Williams PP, Bendel R, Butler J, Farb H, Wager G, Knox EG, Julian T, Thompson TR. Assay of disaturated phosphatidylcholine in amniotic fluid as a test of fetal lung maturity: experience with 2000 analyses. Clin Chem 1987; 33/9: 1648-1651.

26. VanMarter LJ, Berwick DM, Torday J, Frigoletto FD, Wise PH, Epstein MF. Interpretation of indices of fetal pulmonary maturity by gestational age. Pediatr Perinat Epidemiol 1988; 2: 395-399.

VI. APPENDIX

1. Illustration of mathematical correction for verification bias

In this appendix we describe the mathematical correction for verification bias with an example. The basic assumption is that selection for verification (and thus the probability of verification) depends on clinical factors and test results, and does not depend directly on true disease status. An equivalent assumption is that the predictive values are the same for the verified and source population (V.,8).

To understand how the correction for verification works, it is useful to consider a situation where selection for verification depends only on the test result and the test result is dichotomous. For example, consider the L/S ratio using a cutoff value of 2 to distinguish positive from negative tests; that is, an L/S ratio ≤ 2 is an immature test result and an L/S ratio > 2 is a mature test result. Assume, for this example, that the decision to deliver the baby depends only on the test result. The cross tabulation, based on our data set, is given in Table a.

Neglecting the unverified cases we calculate the true positive rate, false positive rate and the predictive values as follows:

true positive rate	32 / 42	= 0.76
false positive rate	106 / 616	= 0.17
predictive value positive	32 / 138	= 0.23
predictive value negative	510 / 520	= 0.98

The probability that the diagnosis is verified (that is the baby is delivered) can be expressed conditionally on the test result as follows:

probability of verification		
given a positive test result	138 / 912	= 0.1513
given a negative test result	520 / 1368	= 0.3801

If a sample is a fraction f of a source population, the source population will be 1/f times the size of the sample. Thus, given the assumption that the probability of verification depends only on the test result, we estimate the source population by dividing each cell frequency by the probability of verification given the test result. The 2 x 2 frequency table of the projected source population is given in table b.

Calculating the true positive rate, false positive rate and the predictive values of the projected source population we get:

true positive rate	211 / 237	= 0.89
false positive rate	701 / 2043	= 0.34
predictive value positive	211 / 912	= 0.23
predictive value negative	1342 / 1368	= 0.98

Table a) Verified sample and number of unverified cases.

	positive	negative
disease	32	10
no disease	106	510
TOTAL VERIFIED	138	520
UNVERIFIED	774	848

Table b) Estimated source population.

	positive	negative
disease	32/0.1513 = 211	10/0.3801 = 26
no disease	106/0.1513 = 701	510/0.3801 = 1342

Therefore, the observed true and false positive rates calculated from the verified sample are lower than the actual rates calculated from the projected source population whilst the predictive values are the same. This example only shows what happens for a 2 x 2 table, but for any categorization of the test results the same arithmetic is applicable. If the test result is divided into k categories and the clinical features into n subgroups, we would correct a 2 x k x n table using the same method. In general we can subdivide the population into cells containing the frequency of a particular combination of clinical features and test results. The probability of verification for a particular cell is the number of verified cases divided by the number of verified plus unverified cases of that cell. The estimated frequency of that cell in the source population is the cell frequency in the verified sample divided by the appropriate probability of verification.

Chapter IX

DISCUSSION

I. DECISION ANALYSIS
1. Structuring the decision
2. Can individual probabilities be based on relative frequencies?
3. Finding appropriate data
4. Assessing patient preferences
5. A decision analysis: a point in time versus a continuing story
6. Decision analysis and clinical practice
II. RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS
1. ROC curves in comparing diagnostic tests
2. Verification bias
3. Uninterpretability
4. ROC curves and daily practice
III. CONCLUSIONS
IV. REFERENCES

The number of decision analyses published in medical journals is gradually increasing. Many clinicians, however, tend to be skeptical about the methodology. There are numerous reasons for the somewhat slow acceptance. Some theoretical aspects of decision theory are poorly understood by clinicians who are not used to thinking in mathematical terms. The validity of the method is questioned by critics, as discussed in section I.2 of this chapter. Besides the theoretical aspects, many practical problems exist in applying decision theory in medicine, discussed in sections I.1,3,4 and 5. Finding appropriate data (I.3) and assessing patient preferences (I.4) are major issues that remain problematic. However, advances are being made with the development and application of decision analytical techniques, computer software and teaching. The first part of this chapter discusses theoretical aspects of decision theory and practical aspects in applying clinical decision analysis. The second part of the chapter discusses theoretical and practical aspects of the use of receiver operating characteristic (ROC) methodology.

I. DECISION ANALYSIS

1. Structuring the decision

Modelling a decision always involves a compromise between simplicity and reality. In modelling a clinical decision one should ensure that the trade-offs involved are represented fairly. Simplifying assumptions should be stated explicitly and the impact of the simplifying assumptions on the outcome of the analysis should be discussed. Even a simple crude decision analysis may solve the greater part of the problem. An illustrative example of a decision analysis was presented in chapter III, which concerned the workup in suspected renovascular hypertension, modelling the trade-off of the risk of a workup versus the benefit of being able to treat a renal artery stenosis, if present.

A decision tree, Markov model or combination of the two are the means of structuring a decision. Choosing which type of model to use may be difficult. If the decision concerns a problem with a short time horizon and a non-recursive structure, then a decision tree provides the more convenient model. In chapter III the choice of workup and treatment in suspected renovascular hypertension is modelled with a decision tree. We chose to keep the time horizon short, namely workup and treatment, and the model does not take into account the possibility of treatment for long-term restenosis after angioplasty. Instead, the model includes the overall long-term clinical success.

If structuring the problem in a decision tree results in multiple branches due to recursive events, modelling the problem, or part of the problem, with a Markov process usually yields a more convenient and computationally quicker model. Markov processes are commonly used to model prognosis of chronic diseases. However, a problem with a short time horizon may also be conveniently modelled as a Markov process if the events are recursive and/or the problem encompasses many time-dependent risks. Chapter IV presents an example of a problem with a short time horizon, modelled with a combination of a Markov process and decision tree structure. The problem concerns the decision whether to intervene for acute urinary tract obstruction and involves time-dependent risks, such as the risk of losing renal function of the affected kidney which depends on the duration of the obstruction. If the decision can be modelled as either a decision tree or (partly) as a Markov process, it helps structuring the problem using more than one method. If the models agree, this supports the results.

To simplify a decision analysis a useful technique is to bias the strategies such that the model underestimates the expected utility of the strategy likely to be best and overestimates the expected utility of the other strategies. The bias may involve the structuring of the problem and/or deciding on which data are appropriate for the analysis. For example, in chapter IV we biased the options away from intervention and towards medical management by assuming that if obstruction of the urinary tract recurs after a drainage procedure, the probability of renal impairment will be determined by the period of time evolved since the initial decision. This implies that the applied

probability of renal impairment is larger than in reality, because obstruction has been temporarily relieved in the meantime. Biasing the options as described was practical in that it helped restrict the number of Markov states. Another bias could be to use lower limits for variables involving risks of expectant management, and upper limits for variables involving risks of interventional procedures. If the analysis shows that a strategy is best, in spite of the bias against it, we can be fairly confident of the result.

2. Can individual probabilities be based on relative frequencies?

The probabilities used in a decision analysis are derived from relative frequency estimates, based on observations of cohorts of patients similar to the case under consideration. The empirical findings are subsequently applied to an individual patient, assuming that what happened on average in the cohort, predicts what can be expected to happen to the individual. Critics of clinical decision analysis question the validity of this assumption, arguing that each patient is a unique individual. However, a clinician uses the same train of thought when implicitly using his/her experience of similar patients to decide on the management of the case at hand. In fact, the above assumption constitutes the cornerstone of the empirical scientific foundation of modern medicine.

For a patient different in some way from the cohort about which data are published, it may be possible to adjust the available data. For example, in chapter IV one of the probabilities used in the model is the probability of haemorrhage as a result of performing a percutaneous nephrostomy. From the literature we estimate this probability to be 0.008. If a patient has an increased bleeding tendency, this probability will be higher. However, exact data on the higher probability are not available. We can adjust the estimated probability of haemorrhage by using a clinician's assessment of the likelihood of haemorrhage based on the severity of the bleeding tendency, as illustrated in chapter IV. At times it may even be necessary to use data from animal experiments to derive probability estimates, an example of which is given in chapter IV.

3. Finding appropriate data

In deriving probability estimates from the literature, a number of issues play a role.

A particular relevant piece of information for a decision analysis may simply never have been determined. Often data exists, but are reported ambiguously or in a cryptic manner. Furthermore, which of the different reported values should be used, or whether a formal meta-analysis should be performed, need not always be evident.

The probability of success or complications of new techniques may be reported as overall experience, or as experience after the learning phase. Again one may question which estimate should be used. The overall experience represents a less optimistic estimate, and probably reflects

the average hospital more reliably. After all, the published reports are usually from clinics with more experience in the reported technique: as has been said "you cannot always be sure that your patient belongs to the last hundred seen in the best clinic." Furthermore, in reporting results of diagnostic tests, unverified cases and uninterpretable test results tend to be disregarded, which may lead to over- or underestimates of the reported test parameters (see sections II.2 and 3).

A related issue is that of publication bias (1,2). Publication bias occurs when study results influence the likelihood of publication. Positive findings or striking results increase the probability that a paper will be written and submitted by the investigators, and the chance that the paper will be accepted by an editor. The chance of publication is greater if the reported results are statistically significant. Statistical variation alone ensures that among the trials of an ineffective agent a number will yield false positive results, ie. will demonstrate a statistically significant effect, and due to publication bias, these false positive results may be published more frequently than trials demonstrating true negative results. Publication bias has consequences for every summary of data and, thus, for decision analysis.

Disease entities, indications for treatment, the nature of complications and the criteria for success are often poorly defined by the authors of a paper, presenting another problem with finding reliable data. If authors do not clearly state their definition of used terminology, their results cannot be generalized to other similar situations.

The denominator problem is a similar issue. Case reports and short series generally provide no information on the relevant denominator. Studies in which a denominator is mentioned, often do not report explicitly which cohort the denominator represents in calculating percentages. This can be very misleading, and give relative frequencies that make no sense if the chosen denominator is not appropriate in the context of the problem analyzed.

Double counting constitutes a related problem. For example, a death may be counted as a percutaneous nephrostomy death, and at the same time as an operative death, which will lead to an overestimate of mortality. Double counting may occur when authors report on the same cases in various publications, without explicitly stating if and when patients have been reported more than once. For example, a patient with a fatal complication may be counted in two series, which may lead to an overestimate of mortality if both series are included in deriving an average mortality. Unfortunately, it is usually impossible to correct the data for double counting and we have to make do with the data there are.

Performing a decision analysis frequently results in questions about probabilities that cannot be answered with a literature review. A hospital information system or an expert in the field may be the only source of information in these circumstances. Hospital information systems can provide a wealth of information but are not always easily accessible. Limitations in retrieval of information from these systems are the quality of the derived information, due to non-uniformity of definitions, and privacy constraints. However, progress is being made and methods are being developed in this respect.

When using probability estimates supplied by experts, we have to keep in mind the limitations of the given estimates. The consulted expert may be subject to bias, especially if he/she requested the analysis. Decision analysts argue that physicians, similar to other human beings, are poor intuitive estimators of probability (34), influenced by all sorts of biases and over-confident of our estimates (see Chapter I). However, this argument cuts in two ways: on the one hand it motivates the use of decision analysis, but on the other hand it also suggests that we should not rely too heavily on expert opinion for probability estimates. If probability estimates given by an expert are the only data available, it is prudent to perform a sensitivity analysis over a wide range of values of the parameter to assess the influence of the parameter.

An advantage of using decision analysis lies in the fact that missing data, essential to the decision, are identified. In this way a decision analysis of the problem can give direction to future research.

4. Assessing patient preferences

Patient preferences are major issues in decision making which are difficult to model explicitly. Although progress is being made, many problems remain unsolved.

In assessing patient preferences with respect to quality of life, one should distinguish between the anticipated quality of life of a hypothetical event as perceived by someone faced with a choice, and the average experienced quality of life of a group of people with morbidity. An analysis for an individual patient incorporates the individual anticipated quality of life, elicited with the methods as described in chapter II. A generic model or cost-effectiveness analysis may require knowledge of the average experience of patients with morbidity, usually determined using questionnaires or interviews and a scoring method. Applying a protocol, based on the average experience of a group of patients, to an individual, may impair the freedom of choice for that individual. In other words, generic decision models and cost-effectiveness analyses may determine which strategy, on average, optimizes management, but the individual retains the right to choose and must make that choice in anticipation of things to come. A decision analysis that incorporates the individual anticipated quality of life, is meant to help the individual make that choice. The unit PALY's (preference-adjusted life years) introduced in chapter II is more meaningful than the unit QALY's (quality-adjusted life years) in distinguishing individual anticipated quality of life from the average experienced quality of life. Furthermore, as explained (chapter II) PALY's can also incorporate a patient's attitude towards risk.

The framing of the questions presents a problem in assessing patient preferences, because framing influences the results (26,35). If an individual chooses the same option irrespective of how the choices are framed, one can be confident the answer reflects the true preference. However, in eliciting a patient's preference one will often find conflicting answers. How to deal with this phenomenon remains, as yet, unresolved.

An associated problem is that the presented questions usually involve some probability of the event occurring. The perceived probability of the event may intermingle with the anticipated value of the outcome, which may subsequently influence the elicited preference.

A third problem is that the available approaches to assess the anticipated quality of life with morbidity give different results (29). To determine the quality of life three basic methods are in use (see chapter II): the direct scaling (or category) method, the time trade-off method and the standard reference gamble. With the direct scaling method one asks the patient to mark his/her anticipated value of the states of ill-health on a linear scale so that the distance between the marks are proportional to the difference between the values of the outcomes. This procedure, although easy to conceptualize, tends to give lower results (in the upper part of the scale) than using the other two methods (29). The time trade-off method directly assesses what length of time in full health the patient perceives as equivalent to a specified period of ill-health. This method, fairly easy to comprehend for most patients, results in lower values than when using the standard reference gamble (29). Decision analysts consider the standard reference gamble the "gold standard" in assessing patient preferences (29). The question posed to the patient is to choose between 1) a certain intermediary outcome or 2) a gamble between a better outcome with probability p and a worse outcome with probability 1-p. The quality factor equals the value of p for which the patient indicates indifference between the certain intermediary outcome and the gamble. The standard reference gamble appears to be difficult to comprehend and it requires more time to elicit a patient's preference.

Another problem in assessing the utilities in the context of a decision analysis is whether the average experience of patients who have experienced the event or morbidity should be taken into account in one way or another. The anticipated valuation of a hypothetical event in the future may be quite different from that when the subject is actually confronted with the event, or the valuation once the event has taken place. A study done to investigate women's attitudes to anaesthesia during childbirth presents a good example of the effect of when the questions are asked (3). Questioned whether they wanted to use anaesthesia during childbirth, the attitudes of the women changed from declining anaesthesia one month before labor, to wanting anaesthesia during labor, to again declining anaesthesia one month after delivery. In mutilating procedures the patient might view the quality of life quite differently once he has actually undergone the procedure and adapted to the resulting way of life. This plays a role, for example, in larynx cancer when deciding between laryngectomy (ie. surgical removal of the larynx), resulting in loss of speech, and radiation of the larynx, which preserves speech but results in shorter survival (21). Clearly, the patient should be asked his/her preference before laryngectomy, because once removed, replacing the larynx is impossible. However, it is not clear-cut whether the attitude and experience of similar patients who have undergone the procedure, should in some way be taken into account. A practical approach would be to inform the patient of the experience of those who have undergone the procedure before assessing his/her own preferences. Apparently, short-term and long-term preferences are not always the same, which may have consequences for the decision.

In eliciting patient preferences the question arises whether it is ethical to confront a patient seriously ill with hypothetical choices. The hypothetical choices are at present the only means of eliciting measures of quality and aversion. The questions posed deal with life, death and morbidity. Most patients do not appreciate being confronted with their illness and possible death. Even healthy subjects are not always willing to respond to these questions.

Summarizing, assessing patient preferences is an important issue in decision making. Quality of life with morbidity and attitudes towards risk are difficult to quantify. These issues are the most problematic and controversial issues in the field of medical decision making.

5. A decision analysis: a point in time versus a continuing story

A published decision analysis gives the situation at one point in time, based on the available data and assumptions at that moment. In fact, a decision analysis is a continuing story, and preferably one should redo the analysis periodically, updating the data and modifying the model to accommodate new concepts, new technology and new results. This applies to the model on renovascular hypertension presented in chapter III. One of the reasons for performing the analysis was that previously published decision analyses on the subject were outdated (19,20,31). McNeil performed an analysis on renovascular hypertension (19,20), comparing medical treatment without diagnostic workup to diagnostic workup followed by surgical treatment should a renal artery stenosis be found. She concluded that to screen all hypertensive Americans with a diastolic blood pressure over 90 mmHg for renal artery stenosis would be far too costly. Diagnostic tests considered were intravenous pyelogram, isotope renogram and/or angiogram. Weinstein and Stason performed a cost-effectiveness analysis of detection and treatment of hypertension (31), concluding that a program designed to improve patient adherence to their medication may be better use of limited resources than wide-scale programs to detect hypertension. Since the time of McNeil and Weinstein's analyses, 1975, technology has changed: a new treatment for renal artery stenosis exists, namely percutaneous angioplasty, and a new diagnostic test has been introduced for renal artery disease, namely intravenous digital subtraction angiography. In addition, operative mortality is currently lower and anti-hypertensive medication has improved. We analyzed the decision of workup and treatment in suspected renovascular hypertension, taking into account these new aspects (chapter III), and focussing on the choice of workup and the choice of treatment. From the analysis we concluded that the prior probability of renal artery stenosis, together with the diastolic blood pressure while the patient is taking antihypertensive medication, determine the choice of workup. In most cases angioplasty is preferable to surgical intervention, except in ostial and occlusive atherosclerotic disease.

However, since we performed the analysis three years ago, new reports on the subject have been published and are being published. Surgery has become safer and the clinical results of surgical intervention are better than three years ago (12,18,25). Angioplasty technology has changed

with safer guidewires and more flexible and thinner catheters, but whether this improves the clinical results must still be determined[1]. The probability of restenosis months to years after angioplasty of a renal artery stenosis, has recently been quantified (7,17). Restenosis after angioplasty appears to be a major problem, occurring in 18% of patients with fibromuscular disease and 43% of patients with atherosclerotic disease (7,17). (Note that restenosis after bypass surgery occurs in about 14% of patients (7,30), often necessitating nephrectomy.) Furthermore, the indications of angioplasty currently include prevention of renal failure due to renal artery stenosis, especially in diabetics. The changing indications of angioplasty may influence the clinical results. Long-term results of angioplasty in improving renal function are currently being determined (17).

A recently published cost-effectiveness study comparing surgery and angioplasty for renal revascularization (7) suggests that surgical intervention should be performed on older patients and patients with atherosclerotic renovascular disease. However, these results are based on a small group of 52 patients and the clinical results of angioplasty incorporated in the analysis are poorer than in previous reports of angioplasty.

Besides the more recent data on angioplasty and surgery for renal artery stenosis, more data are being published on the diagnostic tests involved. In addition, more data are available on the risks of renal failure due to the use of radiographic contrast (23). With the current use of low osmolar contrast medium, the risks of angiography and intravenous digital subtraction angiography (DSA) are possibly lower and the discomfort appears to be less, however, low-osmolar contrast media are three to fifteen times as costly as high-osmolar contrast media (14,16). Furthermore, digital subtraction technique is improving (5): with a reported sensitivity of 100% and specificity of 93%. The value of renal vein renin measurements remains controversial (28).

Angiotensin converting enzyme (ACE) inhibitors have been added to the list of antihypertensive medications (6,8,10). Their use in the treatment of renovascular hypertension is controversial, because renal function may deteriorate (11). However, ACE inhibitors are currently used in the diagnosis of renovascular hypertension. Captopril, an ACE inhibitor, induces changes during radioisotope renography (9) in the presence of renal artery stenosis. Kidneys with renal artery stenosis demonstrate impaired excretion after administration of captopril, which can be explained as follows. A stenosis causes a decrease in perfusion and glomerular filtration rate (GFR), which, by autoregulation, induces an intrarenal increase in angiotensin II to sustain perfusion and filtration. With the administration of an ACE inhibitor this autoregulatory phenomenon is interrupted, and excretion will thus be impaired (9). The specificity of a renogram

---

[1]Note that in our analysis (chapter III) we included major angioplasty related complications (such as rupture of the renal artery) among the angioplasty technical failures, because surgery is necessary for both major procedure related complications and for technical failures. Another simplification was to consider among the medical complications of invasive procedures only permanent major complications that would affect life expectancy, that is myocardial infarction and cerebrovascular accidents. A small renal infarct or transient renal failure was assumed not to affect life expectancy substantially.

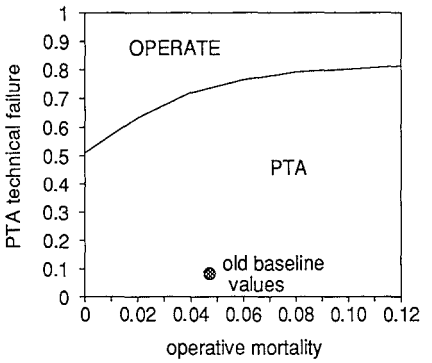
performed after administration of captopril is higher than without captopril (100% compared to 82%) (9 and chapter III). However, the reported sensitivity is lower than that of conventional renography (80 % compared to 86%) (9 and chapter III).

Captopril is also used in a response test: a raised plasma renin level one hour after oral administration of captopril indicates the presence of renovascular hypertension (4,24). The sensitivity of this test lies between 84% and 100% and the specificity between 93% to 95% (4,24). However, in the presence of renal impairment, both sensitivity and specificity are lower (79% and 67% respectively) (24). With the advent of the captopril-renogram and captopril response test, screening to exclude renovascular hypertension with these tests has become less risky compared to performing angiography, and is superior to performing an intravenous pyelogram both in terms of informativeness and in terms of risk. However, should renovascular hypertension be suspected, an imaging procedure will eventually have to be performed if either angioplasty or surgery is considered. (Note that imaging may be part of the angioplasty procedure, as in the "no test-angioplasty" strategy in chapter III.) In this respect our analysis remains valid, except that patients referred to the radiology department will have a higher prior probability of renovascular disease due to more adequate clinical selection.

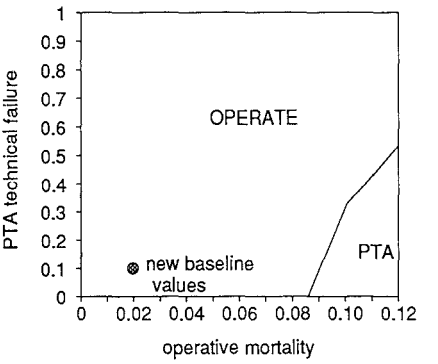
Not only are data constantly being published, but concepts on modelling in decision analysis are changing and advanced software has become available, making more realistic modelling feasible. The most noteworthy progress is the use of Markov processes to calculate prognosis.

We reanalyzed the renovascular hypertension problem after adjusting the model for the new data and concepts discussed above. In particular we took into account renal failure as complication, the captopril test, the captopril renogram, the new data on intravenous DSA, long-term restenosis, and the most recent data on risks and clinical results of angioplasty and surgery. We assume that bypass surgery will be performed if restenosis occurs after angioplasty, and that nephrectomy will be performed if restenosis occurs after surgery. Furthermore, we calculated life expectancy with a Markov process, taking into account the changing mortality rate as a person ages. The excess mortality rate due to hypertension depending on the diastolic blood pressure is now modelled with a polynomial, extrapolating for age with an exponential function. This provides a better fit of the observed data and reflects the changes in mortality rates as a person ages more accurately than the previously used linear regression model. (Even more precise would be to model survival with hypertension using a logistic regression analysis with age, sex and blood pressure as independent variables and survival as dependent variable, using a database currently available in the Netherlands (27).)

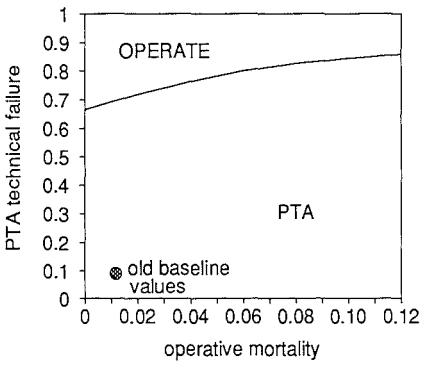
The most pertinent results of the analysis are shown in figures 1 to 3. Figure 1 shows a two-way sensitivity analysis for the operative mortality and the probability of technical failure of angioplasty. The figures of both the previous and revised analysis are shown and the baseline values of the analyzed variables, ie. the values assumed to be applicable, are indicated. Clearly the results have changed: angioplasty is less favorable than before, especially for atherosclerotic



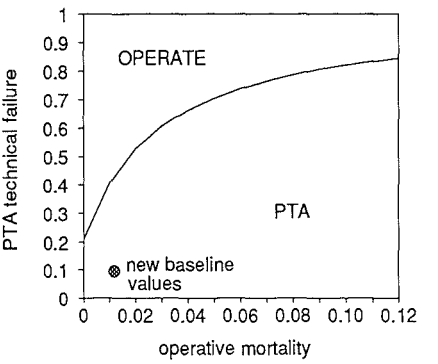
1a) AS, previous analysis



1b) AS, revised analysis



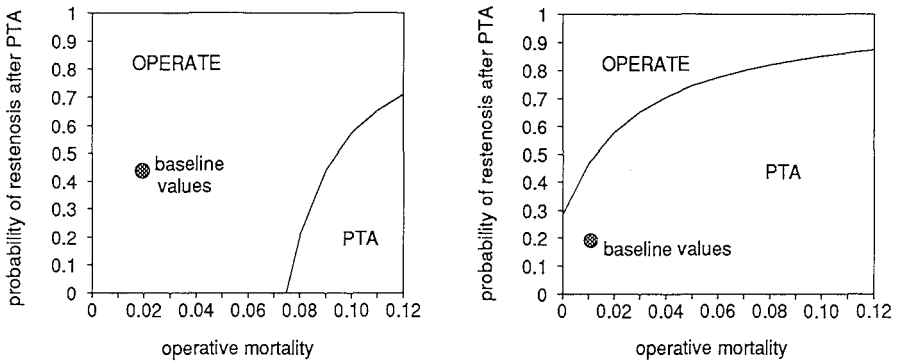
1c) FMD, previous analysis



1d) FMD, revised analysis

Figure 1. Two-way sensitivity analysis for the operative mortality and the probability of technical failure at angioplasty for a) the previous analysis and b) the revised analysis for atherosclerotic (AS) renovascular disease, and c) the previous analysis and d) the revised analysis for fibromuscular dysplasia (FMD). For the area marked "OPERATE", the operative mortality and probability of technical failure at angioplasty are such that operation is preferred. Similarly, for the area marked "PTA", angioplasty is preferred. The baseline values of operative mortality and the probability of technical failure at angioplasty, ie. the values assumed to be applicable, are indicated.

disease. The difference in results can be explained by the high probability of restenosis after angioplasty, the decreased risk of complications with surgery and the improved clinical results of surgery.



2a) AS 2b) FMD

Figure 2. Two-way sensitivity analysis for the operative mortality and the probability of long-term restenosis after angioplasty for a) atherosclerotic renovascular disease (AS) and b) fibromuscular dysplasia (FMD). For the area marked "OPERATE", the operative mortality and probability of restenosis after angioplasty are such that operation is preferred. Similarly, for the area marked "PTA" angioplasty is preferred.

Figure 2 presents the results of a two-way sensitivity analysis for the operative mortality and the probability of restenosis after percutaneous angioplasty. At present, the probability of restenosis constitutes a more controversial piece of information than the probability of technical failure at angioplasty. The results suggest that atherosclerotic disease should be treated with a surgical bypass and fibromuscular dysplasia with angioplasty.

Similar to figure 4 of the previous analysis (see chapter III), figure 3 presents the results of a two-way sensitivity analysis for the diastolic blood pressure with antihypertensive medication and the prior probability of a renal artery stenosis (assuming 63% of stenoses are atherosclerotic). Compared to the previous results, intravenous digital subtraction angiography (DSA) is preferred for a wider range of the prior probability. The option "no test - angioplasty" is optimal only for very high prior probabilities. The captopril test has taken the place of renography, also renography performed after administration of captopril. The results indicate that for a very low prior probability, less than 0.01, and low to intermediate diastolic blood pressure, captopril renography may be preferred as initial test because of the high specificity, claimed to be 100%. However, for such a low prior probability the difference in expected utility of workup and continued medical treatment without workup, is negligible, suggesting that, were we to take monetary costs into account, medical treatment without workup would be preferable.

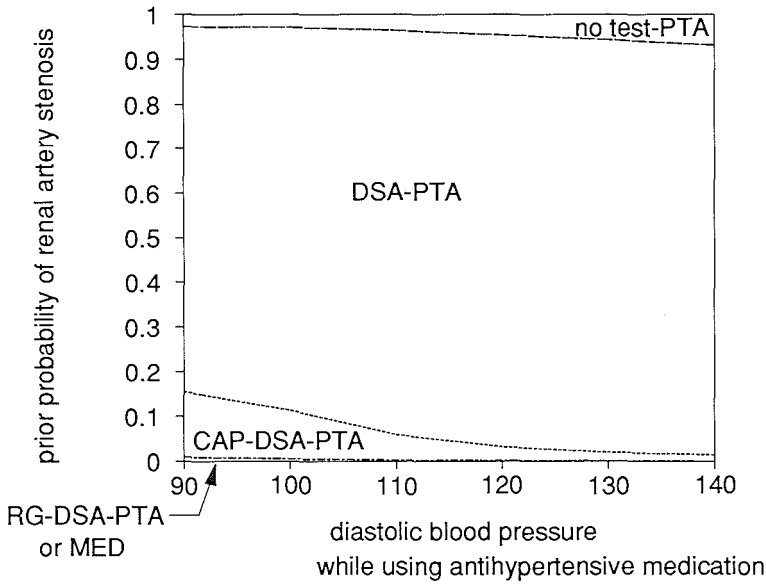


Figure 3. Two-way sensitivity analysis for the diastolic blood pressure with antihypertensive medication and the prior probability of a renal artery stenosis (assuming 63% of stenoses are atherosclerotic). The strategies shown are "continued medical treatment" (MED), "captopril renography - if positive intravenous digital subtraction angiography - if positive angioplasty" (RG-DSA-PTA), "captopril test - if positive intravenous digital subtraction angiography - if positive angioplasty" (CAP-DSA-PTA), "intravenous digital subtraction angiography - if positive angioplasty" (DSA-PTA) and "no test - imaging and angioplasty in the same session" (no test-PTA).

In summary, the results of the revised analysis differs from the previous one mainly in the choice of treatment for atherosclerotic renovascular disease. The results suggest that a surgical bypass should be performed for atherosclerotic stenoses of the renal artery and angioplasty for fibromuscular dysplasia. The lower risk and improved results of modern vascular surgery, together with the recent data on long-term restenosis after angioplasty, have led to this change of results. As far as the workup for renovascular hypertension is concerned, the main conclusion of our previous analysis still holds: the diastolic blood pressure while taking antihypertensive medication, together with the prior probability of a stenosis, determine the choice of workup. The captopril test, with high sensitivity and specificity and negligible risk, serves as a useful screening

test for renovascular hypertension when faced with a low prior probability of renovascular disease. For intermediate and high prior probability, intravenous digital subtraction angiography should be performed initially.

In addition to the idea that an analysis is a continuing story, an analysis may be generic and have to be individualized for a particular patient. Generic models can be used to develop general protocols, however, the decision for an individual patient may require a separate analysis, using data applicable for that particular patient. This may be accomplished by adjusting the risks of intervention, as is done in chapter III for operative mortality if the patient has atherosclerosis and in chapter IV for the percutaneous nephrostomy risks in a patient with an increased bleeding tendency. However, sometimes the analysis has to be redone to model the trade-offs involved for the particular patient.

6. Decision analysis and clinical practice

Even though formal decision analysis is a relatively new field in medicine, the basic concepts and reasoning of decision analysis have been used intuitively by wise and experienced clinicians all along. Traditionally, "clinical judgement" or "clinical intuition" was the process by which medical decisions were made, the process being one of ill-understood hidden logic, pattern recognition and heuristics. In medical school cause-effect type of reasoning is taught, whereas decision making in clinical practice requires reasoning in terms of probabilistic information, which remains a neglected aspect in the training of physicians. While formerly decision making was the domain of wise and experienced clinicians, the introduction of clinical decision analysis gives insight to such decision making, so that it can be critically assessed by all concerned.

Decision analysis may have a major impact on clinical reasoning. Even though a clinician cannot be expected to build a tree for every patient he/she sees, the concepts and reasoning of decision analysis are helpful in clinical practice. Structuring a problem in a logical explicit fashion may give new insights. Merely constructing a decision tree can help a physician understand the pros and cons of the various options. Structuring a problem identifies the information necessary for the decision. Even if some data are imprecise, a crude decision analysis may suggest the optimal management. A decision analytical approach may pinpoint the reason for disagreement, making discussion among decision makers more constructive. Furthermore, it may be helpful in explaining possible management strategies to the patient and family, thereby obtaining informed consent.

In spite of the above mentioned advantages of decision analysis, numerous problems remain when applying the technique, which has consequences for its practical application at present and in the future. Analyzing individual case problems is usually time consuming, so that the decision has often been made before completing the analysis. With increasing experience and

improvement of computer facilities, analyses for individual case problems may become standard clinical practice, but this probably still requires years of research and development. More feasible than decision analyses of individual problems, are analyses of common recurring clinical problems. Such generic decision analyses can help develop clinical guidelines and protocols. Although this is already being done to a certain extent, a major part of decision analytical research in the future could very well focus on the development of guidelines and protocols. This includes guidelines as to the implications of routine test results, whether the report of a chest X-ray or the measured ST depression during exercise testing. Furthermore, with increasing budget constraints, policy makers are showing an interest in cost-effectiveness analyses, the results of which could have a major impact on determining which health programs should have priority.

The major areas of potential technical development in decision analysis include modelling techniques, data retrieval systems, software and the reporting methods. The techniques used to model decisions, probabilistic data and outcome values could be improved by incorporating available methods from mathematics. Access to data retrieval systems, both medical information databases and hospital information systems, must be improved if decision analysis is to become a practical tool in the future. Further development of advanced and user-friendly software will facilitate performing the calculations and may ease experimenting with alternative hypotheses, even for the novice. Improvement of the methods used to report the results of a formal analysis should assist the acceptance of decision analysis, and its recommendations, by clinicians unfamiliar with the technique.

An unresolved issue is whether clinical decision analysis provides better judgement than conventional decision making. Choosing a method to decide which of the two is better confounds matters. Performing a randomized controlled clinical trial in which management in half of the subjects is determined by conventional decision making and in the other half by decision analytical techniques, is far from practical. Decision analysis is not advanced enough to cope with a wide variety of clinical problems. If only one clinical problem is considered the trial would not test the general question posed. Furthermore, what outcome should be considered? For example, suppose conventional decision making suggests medical management, whereas a decision analysis prescribes operation. The patient is operated and dies. Does this prove that conventional decision making is better than decision analysis? The decision analysis took the risk of death into account, and the risk remains present once the decision has been made. Another approach would be to build a decision tree to decide which technique optimizes management. However, using a theory to prove the same theory is obviously invalid (13). In conclusion, deciding whether to use decision analysis or conventional decision making seems more a matter of opinion than a question to be answered by scientific research.

II. RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS

Receiver operating characteristic (ROC) methodology is a method of evaluating and describing test performance and, thus, is important in diagnostic radiology. Radiologists are becoming aware that a likelihood ratio, or equivalently a pair of sensitivity and specificity values, alone does not determine how good a test system performs. The likelihood ratio, sensitivity and specificity depend not only on the performance of the test, but also on the chosen threshold for declaring the test result positive or negative. ROC methodology provides an alternative to the conventionally used parameters. An ROC curve plots pairs of true and false positive rates (see chapter V). The area under the ROC curve provides a measure of overall test performance, an area of 1 representing a perfect test and an area of 0.5 representing a test with no discriminating power.

1. ROC curves in comparing diagnostic tests

Diagnostic tests can conveniently be compared by comparing their ROC curves and specifically the areas under the curves. This was done in chapters VI to VIII for various clinical problems and diagnostic tests. Chapter VI and VII showed that the test performance of promising new technology (in this case MRI) may be less than the high initial expectations, emphasizing that new diagnostic tests should be evaluated rigorously before becoming part of the accepted daily routine. The clinical diagnosis may be considered a diagnostic test and, if appropriate, should be included in the comparison of available tests. For example, in chapter VII we showed that the information obtained by performing CT or MRI in the diagnosis of orbital space-occupying lesions, did not provide incremental information over the clinical diagnosis. Sometimes clinical information does not provide a separate diagnostic test, but may have additional value when interpreting a test. For example, in chapter VIII we showed that using gestational age to predict fetal pulmonary immaturity added information to amniotic fluid tests, while a second test (the SPC test) did not add significant information. Especially with increasing budget constraints, it seems valuable to assess the information provided by clinical evaluation, along with the incremental information provided by expensive new technologies, instead of only comparing imaging and/or biochemical tests. This may lead the way to more efficient use of modern technology.

2. Verification bias

Test parameters and the ROC curve may be biased when not all subjects tested undergo the definitive procedure to determine "true" disease status. The test result, together with clinical

features, usually determine whether a diagnosis is verified or not. If we assume that the probability of verification depends only on the test result and clinical features, but is conditionally independent of true disease status, we can calculate unbiased estimates of the test parameters. The assumption implies that the disease itself only indirectly influences the decision to verify the diagnosis, through its effect on the test result and clinical features. This assumption seems valid when one considers daily practice: we have to decide to proceed with the workup at a time that we know the clinical features and test result, but do not yet know the true disease status. If it is not considered valid, all patients have to undergo the definitive procedure to determine unbiased estimates of the test parameters.

The implication of verification bias for clinical radiology lies mainly in the methods used to evaluate diagnostic tests. Obtaining a "gold standard" for every patient tested is generally considered sound scientific practice. However, this need not always be practical. Patients may be lost to follow-up or it may be unethical to request a gold standard. For example, obtaining histology for all patients who underwent computer tomography and magnetic resonance imaging for a suspected orbital tumor, would have been meticulous (chapter VII) as far as scientific standards are concerned. However, one can hardly expect a patient to undergo enucleation of the orbit, for scientific reasons, when all tests suggest that a malignant lesion is highly unlikely. Numerous other diagnostic tests involve similar problems when being evaluated. Unverified cases of a test under evaluation, unverified for whatever reason, should be reported, and not simply disregarded. This implies that when evaluating a test for its discriminating power for a particular diagnosis, care must be taken to record all patients undergoing the test for that diagnosis. Thus, cases should not be considered immaterial to the series on the grounds that a gold standard is unavailable. Reporting the unverified cases gives an indication of the presence of bias and, if biased, the test parameters and ROC curve can be corrected, provided the assumption as explained above is considered valid. The correction method is fairly simple (see chapter V) and can usually be performed, as long as the unverified cases are not disregarded.

Correcting for verification bias does not necessarily change the ROC curve, even though the individual pairs of true and false positive rates change substantially. If the corrected and uncorrected curves are similar, verification bias shifts the operating points along the curve, but may not substantially effect overall test performance. In such cases, although individual sensitivities and specificities can be seriously biased, the area under the ROC curve provides a measure of test performance less sensitive to bias. We saw this phenomenon in chapter VIII, in which the ROC curves of the amniotic fluid tests were insensitive to verification bias. However, the points on the ROC curve associated with particular cutoff values shifted considerably, and thus, the cutoff value for a particular sensitivity and specificity shifts. Correcting for verification bias, therefore, implies adjusting the chosen cutoff value of the test variable to a different criterion.

3. Uninterpretability

Uninterpretable test results, and unsuccessful procedures, are common occurrences in diagnostic radiology. For example, a patient may have too much bowel gas for a reliable ultrasound or may be unable to retain a barium enema. Sometimes the uninterpretability itself contains diagnostic information. For example, an inaccessible bile tract during transhepatic cholangiography usually indicates an undilated bile tract, implying that extra-hepatic obstruction is unlikely. Inability to retain a barium enema may be due to an obstructive process in the sigmoid. To determine whether an uninterpretable test result contains diagnostic information, we need to determine whether uninterpretability is related to either the disease or absence of the disease. This can be done by calculating the (interval) likelihood ratio of uninterpretable test results, ie. the probability of an uninterpretable test result given disease divided by the probability of an uninterpretable test result given the absence of disease. If uninterpretability is a random phenomenon, ie. it is not related to disease, the (interval) likelihood ratio equals 1. As explained in chapter V, if uninterpretability contains information, uninterpretable test results can be included in the ROC analysis as an additional test result, ranked among the other results according to the (interval) likelihood ratio's.

The clinical implications of uninterpretable test results are 1) uninterpretable results should be reported in publications and 2) an uninterpretable test result may contain diagnostic information. The clinical implication of using uninterpretable test results was shown with the data from the literature on ultrasound for appendiceal disease (chapter V), in which a non-visualized appendix contained information, ie. unlikely to be diseased. In the analysis of amniotic fluid tests for fetal pulmonary maturity (chapter VIII), uninterpretable tests where those in which the amniotic fluid was stained with blood or meconium. The presence of blood was not related to fetal pulmonary maturity or immaturity, ie. blood staining of amniotic fluid is a random phenomenon with respect to disease. This was concluded because the (interval) likelihood ratio of blood stained amniotic fluid tests equals more or less 1. However, the presence of meconium was related to absence of disease, ie. the likelihood ratio was zero. This implies that meconium stained amniotic fluid contains information and suggests that the fetal lungs are mature (equivalent to a negative test result). However, only 25 amniotic fluid samples of the total of 854 verified cases were meconium stained and including meconium stained fluid as an additional test result in the ROC analysis did not influence the results.

Summarizing, physicians should be aware of the potential information contained in an uninterpretable test result and use this information when appropriate.

4. ROC curves and daily practice

ROC curves are mainly useful for comparing diagnostic tests. In daily clinical practice an (interval) likelihood ratio may be more useful than an ROC curve. An (interval) likelihood ratio

is the ratio of the probability of seeing a particular test result in patients with the disease and seeing it in patients without the disease. Using an (interval) likelihood ratio we can adjust the initial clinical impression with the information from the test. Further workup and/or treatment are begun only if the probability of the disease is large enough to warrant further intervention. Given a particular prior probability of disease, the (interval) likelihood ratio above which further workup and treatment are indicated, should be determined not only by how good a test performs, but also by the costs, risks and benefits of true and false positive, and true and false negative test results. For each clinical setting in which the test is used, these costs, risks and benefits should be determined, from which follows the optimal operating point on the ROC curve for that particular clinical problem. From the optimal operating point directly follows the optimal (interval) likelihood ratio, and the cutoff value of the test variable to be used in clinical practice.

ROC methodology is slowly having an impact in diagnostic radiology. Imaging techniques are increasingly being compared using ROC curves and the radiological community accepts the method. Even though numerous articles reporting sensitivity and specificity are still being published, articles reporting ROC curves are accumulating. Methodological issues are also being addressed, both to explain the method and to discuss unresolved methodological problems.

One of the unresolved methodological issues is that, not only should a diagnosis be made, but the location of the process should also be correctly identified. For example, if the diagnosis bronchogenic carcinoma is made, but the reader of the image identifies the right upper lobe as the location while the process is in the left upper lobe, the test result will lead to removal of a healthy part of the lung instead of the diseased part. The LROC, ie. the location receiver operating characteristic curve, also known as the joint ROC, provides a method of including the location in the ROC curve (22,33). In the LROC a true positive test result refers to a test result in which both the diagnosis and the location are correctly identified. A LROC curve is similar to a ROC curve, however, if the reader does not always correctly identify the location, the maximum true positive fraction of the LROC will have a value less than 1. The area under the LROC curve provides a measure of the test performance and can be used to compare tests.

The problem of multiclassification remains one of the major unresolved issues of ROC methodology. An ROC curve as currently used necessitates dividing the final diagnosis in two categories. Usually the division separates the final diagnosis into the presence or absence of disease. The division should have clinical consequences, for example, determining whether a treatment will be instituted. However, in daily radiological practice an imaging test may suggest various diagnoses, together with the radiologist's confidence in each of these diagnoses. The probability of each of the multiple diagnoses, given the test result, and the radiologist's confidence in each, are not independent, but rather interrelated. The performance of a test used to distinguish between multiple diseases can not simply be represented with an ROC curve. A few methods have been developed to construct ROC curves appropriate for multiclassification problems, but these are, as yet, not generally applicable (15,32). Research in this area is a challenge for ROC methodologists.

III. CONCLUSIONS

Decision analysis is an explicit method of making decisions involving uncertainty, in which risks and benefits are quantitatively balanced. Structuring a decision in the form of a decision tree helps identify the trade-offs involved and the information needed. Finding appropriate data for the analysis may be problematic because of difference in used terminology and ambiguous reporting. Assessing patient preferences is a major issue in decision making: preferences are difficult to quantify and the methods used are controversial. The potential impact of decision analysis lies in the method of reasoning, in changing clinical practice and developing protocols, in analyzing complicated individual problems and in cost-effectiveness analysis.

ROC analysis is a method of evaluating and comparing diagnostic tests. ROC analysis can identify what cutoff value of a test variable should be used in clinical practice. Test parameters and ROC curves may be biased if not all subjects tested undergo the definitive procedure to determine "true" disease status. Such unverified cases should be reported and unbiased test parameters and ROC curve may be estimated. Uninterpretable test results may contain diagnostic information if the cause of uninterpretability is related to disease or absence of disease. Uninterpretable test results should be reported and may be used as an extra test result in the ROC analysis. The major impact of ROC analysis lies in the comparison of diagnostic tests.

IV. REFERENCES

1. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. J R Statist Soc 1988; 151: 419-463.

2. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer I 1989; 81:107-115.

3. Christensen-Szalanski JJJ. Discount functions and the measurement of patient's values. Women's decisions during childbirth. Med Decis Making 1984; 4: 47-58.

4. Derkx FHM, Tan-Tjiong HL, Wenting GJ, Man in 't Veld AJ, Seyen AJ van, Schalekamp MADH. Captopril test for diagnosis of renal artery stenosis. In: Glorioso N et al, editors, Renovascular Hypertension. Raven Press, New York, 1987; 295-304.

5. Dunnick NR, Svetkey LP, Cohan RH, et al. Intravenous digital subtraction renal angiography: use in screening for renovascular hypertension. Radiology 1989; 171: 219-222.

6. Edmonds D, Knorr M, Greminger P, Walger P, Frielingsdorf J, Vetter H, Vetter W. ACE inhibitor versus beta-blocker in the treatment of essential hypertension. Nephron 1987; 47: suppl 1: 90-93.

7. England WL, Roberts SD, Grim CE. Surgery or angioplasty for cost-effective renal revascularization? Med Decis Making 1987; 7: 84-91.

8. Franklin SS, Smith RD. A comparison of enalapril plus hydrochlorothiazide with standard triple therapy in renovascular hypertension. Nephron 1986; 44: suppl 1: 73-82.

9. Geyskes GG, Oei HY, Puylaert CBAJ, Mees EJD. Renovascular hypertension indentified by Captopril-induced changes in the renogram. Hypertension 1987; 9: 451-458.

10. Greminger P, Luscher TF, Zuber J, Kuhlmann U, Sxhneider E, Siegenhaler W, Vetter W. Surgery, transluminal dilatation and medical therapy in the management of renovascular hypertension. Nephron 1986; 44: suppl 1: 36-39.

11. Greminger P, Vetter H, Steurer J, Siegenthaler W, Vetter W. Captopril and kidney function in renovascular and essential hypertension. Nephron 1986; 44: suppl 1: 91-95.

12. Grim CE, Yune HY, Donohue JP, Weinberger MH, Dilley R, Klatte EC. Renal vascular hypertension. Surgery vs. dilatation. Nephron 1986; 44: suppl 1: 96-100.

13. Hofstadter DR. Gödel, Escher and Bach: an eternal golden braid. Vintage Books, New York, 1980.

14. Jacobson PD, Rosenquist J. The introduction of low-osmolar contrast agents in radiology. Medical, economic, legal, and public policy issues. JAMA 1988; 260: 1586-1592.

15. Kijewski MF, Swensson RG, Judy PF. Analysis of rating data from multiple-alternative tasks. J Math Psych, in press.

16. Kinnison ML, Powe NR, Steinberg EP. Results of randomized controlled trials of low- versus high-osmolality contrast media. Radiology 1989; 170: 381-389.

17. Kremer Hovinga TK, Jong PE de, Zeeuw D de, Donker AJM, Schuur KH, Hem GK van der. Restenosis prevalence and long-term effects on renal function after percutaneous transluminal renal angioplasty. Nephron 1986; 44: suppl 1: 64-67.

18. Mahler F, Triller J, Weidmann P, Nachbur B. Complications in percutaneous transluminal dilatation of renal arteries. Nephron 1986; 44: suppl 1: 60-63.

19. McNeil BJ, Adelstein SJ. The value of case finding in hypertensive renovascular disease. N Engl J Med 1975; 293: 221-226.

20. McNeil BJ, Varady PD, Burrows BA, Adelstein SJ. Cost-effectiveness calculations in the diagnosis and treatment of hypertensive renovascular disease. N Engl J Med 1975: 293: 216-221.

21. McNeil BJ, Weichselbaum R, Pauker SG. Speech and Survival: Tradeoffs between quality and quantity of life in laryngeal cancer. N Engl J Med 1981; 305; 982-987.

22. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989: 24: 234-245.

23. Moore RD, Steinberg EP, Powe NR, White RI, Brinker JA, Fishman EK, Zinreich SJ, Smith CR. Frequency and determinants of adverse reactions induced by high-osmolality contrast media. Radiology 1989; 170: 727-732.

24. Muller FB, Sealey JE, Case DB, Atlas SA, Pickering TG, Pecker MS, Preibisz JJ, Laragh JH. Acute converting enzyme inhibition with Captopril: a simple screening test for renovascular disease. In: Glorioso N et al, editors, Renovascular Hypertension. Raven Press, New York, 1987; 317-328.

25. Novick AC. Overview. Surgical treatment of renovascular hyertension. In: Glorioso N et al, editors, Renovascular Hypertension. Raven Press, New York, 1987; 447-460.

26. Pauker SG, McNeil BJ. Impact of patient preferences on the selection of theapy. J Chron Dis 1981; 34: 77-86.

27. Personal communication Prof. Dr. J. Lubsen, Center for Clinical Decision Analysis, Erasmus University and Dijkzigt Hospital, Rotterdam.

28. Pickering TG, Sos TA, Vaughan ED, Laragh JH. Differing patterns of renal vein renin secretion in patients with renovascular hypertension, and their role in predicting the response to angioplasty. Nephron 1986; 44: suppl 1: 8-11.

29. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for Health Outcomes: comparison of assessment methods. Med Decis Making 1984; 4: 315-329.

30. Stanley JC. Renovascular hypertension: the surgical point of view. In: Clinical aspects of renovascular hypertension (Schilfgaarde R van, ed). Boston: Martinus Nijhoff 1983; 259-268.

31. Stason WB, Weinstein MC. Allocation of resources to manage hypertension. N Engl J Med 1977; 296: 732-739.

32. Steinbach WR, Richter K. Multiple classification and receiver operating characteristic (ROC) analysis. Med Decis Making 1987; 7: 234-237.

33. Swets JA, Pickett RM. Evaluation of Diagnostic Systems. New York: Academic Press, 1982.

34. Tversky A and Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science 1974; 185: 1124-1131.

35. Tversky A and Kahneman D. The Framing of Decisions and the Psychology of Choice. Science 1981; 211: 453-458.

SUMMARY

Clinical decision analysis is an explicit method of making decisions involving uncertainty about a diagnosis, prognosis, and the risks and benefits of diagnostic tests and/or treatment. Expected utility theory forms the basis of decision analysis (chapter II), the most important assumption being that a person wishes to maximize expected utility or minimize expected disutility. Clinical decision analysis uses decision trees and Markov processes to represent potential clinical strategies, their consequences and outcomes. A simple or non-recursive decision tree represents the potential events with successive branching. A Markov process models different states of health and different transition probabilities for moving from one state to another. The latter model facilitates simulating risks which recur repetitively or change over time. Structuring a decision problem in the form of a model helps identify trade-offs and the necessary critical information. Assessing patient preferences and attitudes towards risks are major issues in decision making, however, preferences and attitudes are difficult to quantify and the methods used are controversial.

An analysis of the optimal evaluation and treatment of suspected renovascular hypertension (chapter III) illustrates the application of decision trees to examine a common problem. Conclusions of the analysis are 1) the choice of workup depends on the prior probability of renovascular disease and the diastolic blood pressure attained while the patient is on antihypertensive medication, and 2) percutaneous transluminal angioplasty is in most cases of renal artery stenosis preferable to surgery or medical treatment. A revision of the analysis (chapter IX) illustrates how the results may change if new data become available. The revised analysis differs from the previous one mainly in the choice of treatment for atherosclerotic renovascular disease. The results suggest that a surgical bypass should be performed for atherosclerotic stenoses of the renal artery and angioplasty for fibromuscular dysplasia. The lower risk and improved results of modern vascular surgery, together with the recent data on long-term restenosis after angioplasty, have led to this change of results. As far as the workup for renovascular hypertension is concerned, the main conclusion of our previous analysis still holds: the diastolic blood pressure attained while the patient is on antihypertensive medication, together with the prior probability of a stenosis, determine the choice of workup. The captopril test, with high sensitivity and specificity and negligible risk, serves as a useful screening test for renovascular hypertension when faced with a low prior probability of renovascular disease. For intermediate and high prior probability of renovascular disease, intravenous digital subtraction angiography should be performed initially.

An analysis of the optimal intervention and the timing of that intervention in patients with urinary tract obstruction caused by urolithiasis (chapter IV) illustrates the usefulness of a Markov process in modelling time-dependent risks. The analysis highlights that percutaneous and retrograde drainage of the urinary tract are both low risk procedures effective in treating urinary tract

obstruction. The model suggests that even for very low likelihoods of pyonephrosis, the urinary tract should be drained to prevent renal impairment and death from sepsis. However, in patients with urinary tract obstruction without signs of pyonephrosis, postponing intervention between 5 to 15 days may be a reasonable alternative, giving the stone an opportunity to pass spontaneously.

Receiver operating characteristic (ROC) methodology evaluates and describes test performance and, thus, is important in diagnostic radiology (chapter V). An ROC curve plots the true versus the false positive rates, with the area under the curve providing a measure of overall test performance. Test parameters and the ROC curve may be biased if not all subjects tested undergo the "gold standard" procedure to determine "true" disease status, which is called verification bias. If we assume that the probability of verification depends on the test results and clinical features, but is conditionally independent of true disease status, we can calculate unbiased estimates of the test parameters and ROC curve. Uninterpretable test results may contain diagnostic information if the cause of uninterpretability is related to either disease or absence of disease, in which case uninterpretability may be used as an additional test result in the ROC analysis.

An ROC analysis of the performance of computer tomography and magnetic resonance imaging in the assessment of extension of neoplastic disease of the nasopharynx, nose and paranasal sinuses and the parapharyngeal space (chapter VI), demonstrates no statistical significant difference in overall test performance. However, magnetic resonance imaging is superior to computer tomography in evaluating regions involving predominantly soft tissue structures and comparatively large bony structures, whereas computer tomography performs better than magnetic resonance imaging in evaluating regions involving thin bony structures.

An ROC analysis of the clinical diagnosis and the diagnosis made on computer tomography and magnetic resonance imaging of orbital space-occupying lesions (chapter VII) shows that, after correcting for verification bias, magnetic resonance imaging performs slightly better than computer tomography. However, neither magnetic resonance imaging nor computer tomography provide significant incremental information over and above that of the clinical evaluation in the diagnosis of orbital tumors. It should be pointed out that the group studied was limited in size and that all patients could be examined very well by means of ophthalmoscopy.

The lecithin/sphingomyelin (L/S) ratio and the saturated phosphatidyl choline (SPC) measurement are amniotic fluid tests used to assess fetal pulmonary maturity and predict the development of respiratory distress syndrome of the newborn. The L/S and SPC tests are evaluated with an ROC analysis (chapter VIII) involving the effect of covariates, verification bias and combination testing. The effect of covariates on the ROC curves is analyzed with a regression methodology, which includes all available data in the construction of an ROC curve for each subgroup. We used a logistic regression analysis to model the probability of verification, which facilitates correcting for verification bias of a fully stratified data set in spite of small cell frequencies. We examine combination testing with prediction rules using prospective logistic modelling, with the test results and clinical features as independent variables and the probability

of respiratory distress syndrome as dependent variable. Our analysis indicates that the L/S ratio is a significantly better test than the SPC test, suggesting that the L/S ratio should be used in clinical practice. For high gestational age the L/S and SPC tests perform better than for low gestational age. Contamination of the specimen, with blood and/or meconium, degrades the ROC curves. Correcting for verification bias does not influence the ROC curves significantly but changes the cutoff value of the test variable for any particular operating point substantially. The use of prediction rules to evaluate combination testing shows that performing the SPC test in addition to the L/S test adds no significant information compared to performing only the L/S test. Including gestational age in the prediction rule of either test improves the prediction significantly.

In conclusion, decision analysis explicitly models the risks and benefits of alternative clinical practices and may help in determining the optimal management of clinical problems. The major impact of ROC methodology lies in evaluating and comparing diagnostic tests. Correcting for verification bias and including uninterpretable test results in the ROC analysis may influence the results.

NEDERLANDSE SAMENVATTING

Klinische besliskunde is een expliciete methode ten dienste van de besluitvorming in de geneeskunde. Hierbij worden betrokken de onzekerheid over een gestelde diagnose, de consequenties van een potentiële ziekte, de risico's en baten van diagnostisch onderzoek en/of die van behandeling. De belangrijkste onderliggende veronderstelling van besliskunde is dat een mens kiest voor de optie met de grootst mogelijke verwachte waarde. In de klinische besliskunde (hoofdstuk II) worden beslisbomen en Markov processen gebruikt om potentiële strategieën, hun consequenties en de uitkomsten te representeren. Een beslisboom representeert mogelijke hypothetische gebeurtenissen in de toekomst door middel van successievelijke vertakkingen. Een Markov proces modelleert de verschillende gezondheidstoestanden en de overgangen van de een naar de ander, wat goed bruikbaar is als een risico zich herhaalt of verandert in de tijd. Een beslissing structureren in de vorm van een model identificeert de overwegingen van de beslissing en de benodigde informatie. Het vaststellen van de voorkeuren van patiënten en attituden jegens risico's zijn belangrijke aspecten van klinische besluitvorming, echter, voorkeuren en attituden zijn moeilijk te quantificeren en de gebruikte methoden zijn controversieel.

Een model betreffende diagnostisch onderzoek en behandeling van hypertensie vermoedelijk veroorzaakt door een nierarteriestenose (hoofdstuk III) illustreert de toepassing van beslisbomen om een veel voorkomend probleem te analyseren. Conclusies van de analyse zijn 1) de keuze van diagnostisch onderzoek hangt af van de a priori kans op een nierarteriestenose en de bereikte diastolische bloeddruk bij gebruik van antihypertensiva, en 2) percutane transluminale angioplastiek (Dotteren) is voor de meeste patiënten met een nierarteriestenose te verkiezen boven operatie of medische behandeling. Een herziene versie van de analyse (hoofdstuk IX) illustreert hoe de resultaten kunnen veranderen als nieuwe informatie beschikbaar komt. De herziene analyse verschilt met de vorige voornamelijk wat betreft de keuze van behandeling voor atherosclerotische nierarteriestenosen. De resultaten suggereren dat een atherosclerotische nierarteriestenose chirurgisch behandeld dient te worden terwijl voor fibromusculaire dysplasie angioplastiek de beste keuze is. De lage risico en verbeterde resultaten van moderne vaatchirurgie, samen met gegevens over restenosering na angioplastiek, hebben geleid tot deze veranderde inzichten. Wat betreft de keuze van diagnostiek blijft de voornaamste conclusie van de vorige analyse geldig: de keuze hangt af van de a priori kans op een nierarteriestenose en de bereikte diastolische bloeddruk bij gebruik van antihypertensiva. De captopril test, met hoge sensitiviteit en specificiteit en verwaarloosbare risico's, is een goede screenings test voor renovasculaire hypertensie als de a priori kans op een nierarteriestenose klein is. Voor een intermediaire of hoge kans op een nierarteriestenose is intraveneuze digitale subtractie angiografie te verkiezen als initieel diagnostisch onderzoek.

Een analyse betreffende de vraag wanneer het geïndiceerd is in te grijpen voor urineweg obstructie veroorzaakt door een uretersteen (hoofdstuk IV) illustreert de bruikbaarheid van een Markov proces voor het modelleren van tijds-afhankelijke risico's. De analyse benadrukt dat percutaan en retrograad draineren van de urineweg ingrepen zijn met weinig risico en effectief in de behandeling van een geobstrueerde urineweg. Als er mogelijk infectie van de urineweg aanwezig is, moet de urineweg gedraineerd worden om functie verlies van de nier en fataal verloop van de infectie te voorkomen. Een patiënt met een geobstrueerde urineweg zonder tekenen van infectie kan 5 tot 15 dagen geobserveerd worden voordat wordt ingegrepen, om de patiënt een kans te geven de steen spontaan te lozen.

Receiver operating characteristic (ROC) analyse (hoofdstuk V) is een methode om diagnostische testen te evalueren en te vergelijken, en dus belangrijk in de radiodiagnostiek. Een ROC curve is een grafiek van de kans op een positieve test gegeven de ziekte, uitgezet als functie van de kans op een positieve test gegeven afwezigheid van de ziekte. Het oppervlak onder de curve is een maat voor het discriminerend vermogen van de test. Testparameters en de ROC curve kunnen vertekend zijn als niet alle onderzochte patiënten ook de "gouden standaard" procedure ondergaan om de "ware" toestand (wel of niet ziek) vast te stellen. Dit heet verificatie vertekening. Als wij aannemen dat de beslissing om te verifiëren wordt bepaald door het resultaat van diagnostische testen en de klinische gegevens, echter conditioneel onafhankelijk is van de ware ziektetoestand, dan kunnen wij zuivere schattingen berekenen van de testparameters en ROC curve. Testresultaten die niet interpreteerbaar zijn kunnen diagnostische informatie bevatten als de oorzaak hiervoor gerelateerd is aan de ziektetoestand, en kunnen opgenomen worden als een additioneel testresultaat in de ROC analyse.

Een ROC analyse van de beoordeling van doorgroei van tumoren in het gebied van nasopharynx, neus en paranasale sinussen en de parapharyngeale ruimte door computer tomografie en magnetische resonantie (hoofdstuk VI) laat, over het geheel genomen, geen statistisch significant verschil zien tussen de twee beeldvormende technieken. Echter, gebieden met voornamelijk weke delen en relatief grote bot structuren zijn beter te beoordelen met magnetische resonantie dan met computer tomografie, terwijl gebieden met fijne benige structuren beter met computer tomografie zijn te beoordelen.

Een ROC analyse van de klinische diagnose en de diagnoses gesteld met computer tomografie en magnetische resonantie van ruimte innemende processen van de orbita (hoofdstuk VII) laat zien dat, na correctie voor verificatie vertekening, magnetische resonantie beter is dan computer tomografie. Echter, er kan niet worden aangetoond dat computer tomografie of magnetische resonantie een statistisch significante hoeveelheid informatie bijdraagt aan de klinische diagnostiek van orbita tumoren. Er moet worden opgemerkt dat deze studie een vrij kleine groep patiënten betrof die allen goed waren te onderzoeken met oogspiegelen.

Diagnostische testen op vruchtwater (de lecithine/sphingomyeline (L/S) ratio en gesatureerde phosphatidyl choline (SPC) meting) worden gebruikt om de rijpheid van de foetale longen te bepalen en het ontwikkelen van respiratory distress syndroom van de pasgeborene te voorspellen.

Een ROC analyse om deze testen te vergelijken (hoofdstuk VIII) brengt een aantal methodologische problemen met zich mee, namelijk het effect van covariabelen, verificatie vertekening, en de waarde van combinaties van testen. Het effect van covariabelen op de ROC curve werd geanalyseerd met behulp van een regressie techniek, welke rekening houdt met alle beschikbare informatie bij het berekenen van een ROC curve van een deelgroep. Om te corrigeren voor verificatie vertekening werd een logistische regressie techniek gebruikt om de kans op verificatie te modelleren, waarbij correctie mogelijk is voor een volledig gestratificeerd gegevensbestand ondanks kleine celfrequenties. Om combinaties van testen te evalueren zijn er eerst voorspellingsregels afgeleid, met de testresultaten en klinische gegevens als onafhankelijke variabelen en de kans op respiratory distress syndroom als afhankelijke variabel. Vervolgens hebben wij de extra informatie berekend die verkregen wordt uit het doen van een additionele test, gebruik makend van regressie technieken en de bijbehorende statistische testen. De analyse laat zien dat de L/S ratio significant beter is dan de SPC test zodat het aan te bevelen is de L/S ratio te gebruiken. Beide L/S en SPC testen doen het beter voor lange dan voor korte zwangerschapsduur. Verontreiniging van het vruchtwatermonster met bloed en/of meconium vermindert de betrouwbaarheid van beide testen. Correctie voor verificatie vertekening heeft geen significant effect op de ROC curves maar verandert wel, voor een gegeven punt op de ROC curve, de grenswaarde van de test grootheid waaronder de diagnose wordt gesteld. Het gebruik van voorspellingsregels om combinaties van testen te evalueren laat zien dat de SPC test geen informatie bijdraagt aan de L/S ratio en dat het derhalve zinloos is om de SPC test uit te voeren. Het opnemen van de zwangerschapsduur in de voorspellingsregel van een vruchtwaterbepaling verbetert de nauwkeurigheid van de voorspelling.

Concluderend, besliskunde kan een bijdrage leveren aan het klinisch redeneren, kan helpen bij de evaluatie van de klinische praktijkvoering en kan helpen bij de besluitvorming voor ingewikkelde individuele klinische problemen. De belangrijkste bijdrage van ROC methodologie ligt in het evalueren en vergelijken van diagnostische testen. Correctie voor verificatie vertekening en het opnemen van niet interpreteerbare testresultaten in de ROC analyse kan de resultaten beïnvloeden.

CURRICULUM VITAE

Maria Hunink, born March 24, 1958, studied mathematics and medicine at the University of Leiden. She passed the "kandidaats" examination in applied mathematics in 1980, the physician qualifying examination (cum laude) in 1984 and the Foreign Medical Graduates Examination in the Medical Sciences in 1988. From November 1984 to October 1988 she did her residency training in diagnostic radiology at the Free University in Amsterdam and successfully completed the Examination of the Dutch Society for Radiodiagnosis. Since 1983 she has worked on clinical decision analysis and has written various articles and given presentations on the subject. As of November 1988 she works for the Center for Clinical Decision Analysis at the Erasmus University and Dijkzigt Hospital in Rotterdam. Part of this thesis was written during a visiting research fellowship at the Division of Clinical Decision Making, New England Medical Center, Tufts University and the Department of Radiology, Brigham and Women's Hospital, Harvard Medical School in Boston.

ABBREVIATIONS

$P(T^+ \mid D^+)$	the probability of a positive test result given disease synonyms: sensitivity, true positive fraction, true positive rate
$P(T^+ \mid D^-)$	the probability of a positive test result given absence of disease synonyms: 1-specificity, false positive fraction, false positive rate
$P(D^+ \mid T^+)$	the probability of disease given a positive test result synonym: predictive value (positive) of a positive test result
$P(D^+ \mid T^-)$	the probability of disease given a negative test result synonym: predictive value (positive) of a negative test result
AG	angiography
CT	computer tomography
DSA	digital subtraction angiography
ENT	ear, nose and throat
ESWL	extracorporeal shockwave lithotripsy
EU	expected utility
interval LR	interval likelihood ratio, that is the marginal probability of a positive test result given disease divided by the marginal probability of a positive test result given absence of disease marginal true positive fraction divided by the marginal false positive fraction
iv	intravenous
IVU	intravenous urography
LE	life expectancy
LR	likelihood ratio, that is the probability of a positive test result given disease divided by the probability of a positive test result given absence of disease true positive fraction divided by the false positive fraction
L/S	lecithin / sphingomyelin
MED	medical treatment
MRI	magnetic resonance imaging
OPER	surgical treatment, operation
PALY	preference-adjusted life year
PCN	percutaneous nephrostomy
PCNL	percutaneous nephrolithotomy

PTA	percutaneous transluminal angioplasty
QALY	quality-adjusted life year
RDS	respiratory distress syndrome of the newborn
RG	renography
ROC	receiver operating characteristic
RUS	retrograde ureteral stenting
SOL	space-occupying lesions
SPC	saturated phosphatidyl choline
URS	ureteroscopic stone manipulation
US	ultrasound

INDEX

ACE inhibitors, 202
algorithm, 6
angiography, 48
angioplasty, 48, 202
  restenosis after, 203
artificial intelligence, 7
assumptions, 15
attitudes towards risk, 18, 37
averaging out, 5, 19, 93

baseline value, 93
Bayes theorem, 114
bias, 3, 175, 189, 196
  publication, 198
  uninterpretability, 136, 189, 198
  verification, 125, 175, 198, 209
  workup, 125
biostatistics, 6

captopril, 202
  response test, 203
category, 122
category method, 34, 200
cell frequency, 131
certainty effect, 43
certainty equivalent, 38
clinical epidemiology, 6
computer tomography, 140, 154
confidence threshold, 119, 156
contrast
  adverse reaction, 54, 202
  costs, 202
cost-benefit analysis, 8
cost-effectiveness analysis, 8
covariate, 174, 176
cumulative failure, 24, 25, 70
cumulative incidence, 24, 25, 70
cutoff value, 118

DEALE-method, 29, 55, 63, 89
decision tree, 14, 48, 196
denominator problem, 198
descriptive model of decision making, 42, 43
dichotomous finding, 115
dichotomous test, 115
digital subtraction angiography, 202, 48
direct scaling method, 34, 200
disutility, 5
dominance, 17
Dorfman and Alf, 122
DSA, 48

expected utility, 5, 19, 42, 57, 93
expert systems, 7, 198

failure function, 24, 25
false positive rate, 114
flow chart, 6
folding back, 5, 19, 57, 64, 93

generic decision models, 8
Gompertz, 33

Hanley-McNeil algorithm, 122
hazard function, 24, 26, 70, 91
hazard rate, 24, 26, 70, 91
heuristics, 2
  anchoring and adjustment, 2
  availability, 2
  representativeness, 2
hypertension
  renovascular, 48, 201

incidence density, 24, 26, 91
interval likelihood ratio, 115, 136, 211
intuitive decision making, 4, 207
isolation effect, 43

lecithin\sphingomyelin ratio, 174
life expectancy, 18, 22, 30, 55, 203
likelihood ratio, 114, 209
  and verification bias, 132

magnetic resonance imaging, 140, 154
  signal intensities, 140, 154
Mann-Whitney, 122
  and standard error of the area, 122
Markov
  chain, 24
  cycle length, 23, 29, 69, 89
  homogeneous, 24
  memory matrix, 92
  model, 22, 28, 69, 89, 106, 196, 203
  process, 22, 28, 69, 89, 106, 196, 203
  semi-Markov process, 92
  state, 22, 89
Markovian assumption, 23
maximum likelihood estimation, 123
mortality rate
  average, 30, 55, 63
  excess, 30, 31, 55, 63, 80, 85, 203
  patient specific, 30, 55, 63
multi-attribute scale, 5

node
  Boolean, 15
  chance, 15
  decision, 15
  label, 15
  terminal, 15

odds, 114
outcome, 17
outcome value, 17

percutaneous nephrostomy, 66, 76
  complications, 77
  efficacy, 77
  mortality, 77
  successful placement, 76
phosphatidyl choline, 174
prediction rule, 61, 175, 178
  logistic regression analysis, 178
predictive value, 118, 128
preference-adjusted life year (PALY), 34, 41, 199
preferences, 18, 34, 199
prevalence, 24
prior probability, 2, 24, 49, 59
probability, 2, 17, 24, 114, 197
  time-dependent, 28, 196
probability density function, 26
probability of verification, 128
  logistic regression analysis, 178
prospect theory, 43
protocol, 6
PTA, 48
pyonephrosis
  ultrasound, 76

quality-adjusted life year (QALY), 34, 199
quality factor, 34
quality of life, 18, 34, 199

rate, 24
rating method, 121
receiver operating characteristic curve,
  and likelihood ratio, 116
  and rating scale, 118, 144, 158
  and sensitivity, 117
  and specificity, 117
  comparing tests, 123, 140, 145, 154, 158, 179, 209

computer tomography, 140
Dorfman and Alf, 122, 179
Hanley-McNeil algorithm, 122, 179
indices, 122
joint, 212
location, 212
magnetic resonance imaging, 140
Mann-Whitney, 122, 145, 179
maximum likelihood estimation, 123
multiclassification, 212
nonparametric estimate, 122
optimal operating point, 123
parametric estimate, 122
reflection effect, 42
relative frequency, 2, 24, 197
renal artery stenosis, 48, 202
renography, 48
with captopril, 203
renovascular hypertension, 47, 201
respiratory distress syndrome, 174
retrograde ureteral stenting, 66, 78
complications, 79
efficacy, 79
mortality, 79
successful placement, 78
risk aversion, 18, 37
risk neutral, 39, 55
risk seeking, 18, 37
ROC, 113
Rosser and Kind rating scale, 35

sensitivity, 117
sensitivity analysis, 19, 57, 94, 203
single-attribute scale, 5
space-occupying lesion, 140, 154
specificity, 117
standard reference gamble, 35, 200
strategy, 14, 15

technology assessment, 7
threshold, 20, 57
threshold for the diagnosis, 118, 122, 156
time horizon, 17, 68, 89, 196
time trade-off method, 34, 200
trade-off, 15, 196
transition probability, 23, 28, 91
true positive rate, 114

ultrasound
pyonephrosis, 76
urinary tract obstruction, 76, 97
ultrasound for appendiceal disease
and uninterpretable test results, 136
and verification bias, 126
uninterpretability bias, 136
uninterpretable test results, 136, 211
urinary tract obstruction, 66
percutaneous nephrostomy, 76
retrograde ureteral stenting, 78
surgical intervention, 80
ultrasound, 76
urography, 48
urolithiasis, 66
extracorporeal shockwave lithotripsy, 81
percutaneous nephrolithotomy, 81
spontaneous passage, 69
ureterolithotomy, 81
ureteroscopic stone manipulation, 81
utility, 5, 17, 55, 123
utility curve, 37
intern's, 41

verification bias, 125, 151, 175, 209
and likelihood ratio, 132
correction for, 125, 158, 177, 209
logistic regression analysis, 178
verification ratio, 134

Weibull, 33, 70
workup bias, 125

krips repro meppel