CONTRIBUTIONS TO A RATIONAL DIAGNOSIS AND TREATMENT OF LUMBAR DISK HERNIATION

CONTRIBUTIONS TO A RATIONAL DIAGNOSIS AND TREATMENT OF LUMBAR DISK HERNIATION

Bijdragen aan een rationele diagnose en therapie van lumbale hernia nuclei pulposi

Proefschrift

ter verkrijging van de graad van Doctor
aan de Erasmus Universiteit Rotterdam

op gezag van de Rector Magnificus

Prof. Dr. C.J. Rijnvos

en volgens besluit van het College van Dekanen.

De openbare verdediging zal plaatsvinden op
woensdag 28 maart 1990 om 13.45 uur

door

JAN WILLEM PIETER FRANCISCUS KARDAUN

geboren te 's-Gravenhage

Promotiecommissie

Promotor:	Prof. Dr. R. Braakman
Overige leden:	Prof. Dr. Ir. J. H. van Bemmel
	Prof. Dr. H.E. Schütte
	Prof. Dr. A. Staal

"Denn man sieht nur die im Lichten
Die im Dunkeln sieht man nicht."

Für DD

CIP-gegevens Koninklijke Bibliotheek, Den Haag

Kardaun, Jan Willem Pieter Franciscus

Contributions to a rational diagnosis and treatment of lumbar disk herniation / Jan Willem Pieter Franciscus Kardaun. — Delft : Eburon.
Proefschrift Rotterdam. — Met lit. opg. — Met samenvatting in het Nederlands.
ISBN 90-5166-139-8
SISO 604.3, UDC 616.832-07/-08(043.3)
Trefw.: Rugklachten

Typesetting: J.W.P.F. Kardaun
Cover design: W. Smetek
Printed by: Haveka, Alblasserdam

Published by:
Uitgeverij Eburon
Postbus 2867
2601 CW Delft, The Netherlands

ISBN 90-5166-139-8

CONTENTS

INTRODUCTION

Diagnosis and treatment of low back pain has always been an intriguing area. Even if the attention is focussed on the lumbar radicular syndrome (LRS), hereby neglecting other pathology that might co-exist in the lower part of the spine, many professional views are expressed that are at least partially contradictory. If the main cause of a lumbar radicular syndrome, the lumbar herniated disk (LDH), is for a moment considered together with an LRS, then it is said that apparently:

- an LDH is often asymptomatic[11, 21] (implying that it should usually not be treated at all).
- an LDH can show spontaneous regression[31]
- acute LRS episodes run a relatively short course in most cases, regardless of the treatment administered[27]
- an LDH/LRS has to be treated in order to end the "present bout of complaints" [18] (implying that treatment gives often temporary relief)
- any LDH encountered should be treated[37]
- a patient with low back pain and/or LRS should not be automatically referred to the surgeon[36]
- the preferred treatment is nucleotomy[5, 7]
- the preferred treatment is chemonucleolysis (CNL)[8, 13, 22, 30]
- CNL is a quite safe procedure[1, 19, 29]
- CNL is an irresponsably unsafe procedure[28, 35]
- CNL induced chemical changes last only a few months[4]
- computer tomography (CT) is to be preferred over myelography for the radiological diagnosis of LDH[3]
- myelography is to be preferred over CT.[16]

Aim

There is a possibility that some of the above mentioned statements that seem contradictory at first glance will perhaps hold true in peaceful co-existence at second glance. Yet it will be undeniable that some confusion will be part of the researcher or clinician, interested patient, health care planner,[9] medical auditor[2] or journal editor,[15] who wants to know what is the prevailing professional view and the *state of the art* of the diagnosis and therapy of this disease. All of these — except for the researcher — do not have enough time, funds and energy to spend several

years in collecting and analysing data for their question and fill in possible white areas in our map of knowledge. Most likely they will take refuge in reviewing existing data from the literature, and will try to sift and sort the differently qualified evidence and to combine equivalent data and weigh it against opposite facts.

Well, this (and nothing more) is the purpose of the present thesis. Be it that an attempt has been made towards a rational and rationalising way of analysis and synthesis of the research question. This approach, in the past decades borrowed from the military and the econometric and the psychometric [10, 14, 17, 23, 24, 26, 32-34] domains, can be applied to medical problems as well and is then called "medical decision analysis" (in Dutch *medische besliskunde*). The result is not an accurate description of all aspects of diagnosis and treatment of LDH, but rather a crude representation of the major choices that have to be made in the course of "managing a patient" with this disease. Advocates of this method, emphasizing that decision makers cannot wait and have to adhere to *existing* data sources, claim that the required data usually can be found in the literature. This claim has waned, however.[14] Another movement, that of *meta-analysis*[6] is indeed exploiting the literature, but indicates that the gold is buried very deep.

Given the disagreement that exists on the management of LDH, as indicated above, it will be understandable that the literature cannot be unified completely. A "declared" advantage[38] of a decision-analytic representation of a problem is that it is explicit, and allows every reader to disagree *on specific points*, rather than in general. Clinical decision analysis, which started out quite straightforward,[38] nowadays features sophisticated methods (see the current issues of the journal *Medical Decision Making*). Our research problem could be made more elaborate and complicated, ad infinitum, than in chapter 6. This would not lead to more reliable and useful answers.

Given the unstructured way in which data about diagnosis and therapeutic aspects of LDH are usually produced and the limitations in size and representativity of any single-center study, it has to be accepted that some parameters will remain a mystery. Examples are complication rates for the radiological procedures involved, or the total number of days spent in the hospital because of an LDH by each patient. Some other parameters, like the outcome (or result) of the treatment, are only fragmentarily known. A combined assessment of aspects of improvement and the influence of the course of time has to be crudely estimated and artificially constructed. For the studies in the following chapters a considerable effort has been made to collect available data on the major parameters of the research question. In fact the emphasis is more on assessing these parameters, than on designing a new or comprehensive or intellectually satisfying model of diagnosis and treatment of LDH. There are still many open ends, which should be, rather than a disappointment, an encouragement to the knowledgeable reader to

pick up one of these loose ends. An abundance of data on low back pain and lumbar disk herniation is available. Attempts to structure them will be most rewarding.

Roadmap

The major streams of current interest, in the mid-eighties, in the management of a patient with a LRS consider the issues

(a) what is the radiological investigation-of-choice and

(b) is chemonucleolysis a good alternative for nucleotomy for certain patients.

Hereby the history and the physical examination are slightly neglected, but there have not been major developments in these during the last two decades. The advent of a new radiological technique (i.c. computed tomography, CT) is usually an incentive to evaluate its merits. (E.g. chapter 1.) Unfortunately, a thorough evaluation can only take place when the new technique is reasonably mature and sufficiently wide-spread and has been published about in some quantity (chapter 3). Moreover, to evaluate a new technique, reference data for the older technique are needed as well. In this case, these were not immediately available, and had to be collected and analysed as yet (chapters 2 and 3). The results indicate that it is sometimes very worthwhile to evaluate established techniques. Due to several co-determinants of the results of comparative studies (chapter 2), the conclusions of single studies which compare CT and myelography directly (e.g. chapter 1) are usually not generalizable (chapter 3) and other ways have to be sought to combine such studies with other studies describing single radiological investigations (chapter 3).

The other major issue, the relative merits of chemonucleolysis, a "new" therapy which has been *in* and *ex gratia* alternatingly since its announcement 25 years ago, as compared with the established therapies, i.e. bedrest with other non-surgical therapies and with surgical removal of the nucleus pulposus, presented a quite analogous situation. The new therapy was better studied than the established ones. Especially regarding complications of surgical treatment, there were only scattered and isolated data (chapter 4). Mostly, they were side-products of other investigations, in spite of early publications that reported a mortality of approximately 3 per thousand[20, 27, 39] and an infection rate of at least 1%.[25, 27]

In chapter 5, an attempt has been made to fill in this gap, even though it is realized that there probably will be an underreporting of complications in the hospital discharge records.

Integration of the new facts and insights on the radiodiagnosis and the treatment of LDH (chapter 6) has as its most visible effect that it reveals the as yet underdeveloped areas. In the management of low back pain and of LDH such an area is undoubtedly the assessment of the degree of impairment and consequently the

degree of improvement.[12] But even with the presently available data, conclusions can be drawn (chapter 6), if a considerable simplification of outcome assessment is applied.

Questions & Answers

Some of the questions that will be touched upon in the following chapters, and that will be hopefully brought closer to an answer, are formulated here *a posteriori*. Some of the questions are the explicit subject of one of the following chapters, other questions are treated more implicitly. For these, I hope that this enumeration will serve as an appetizer to encourage the reader to make, with help of this publication, a critical assessment of the answers himself.

Questions on radiology:

- Is CT or myelography "better" and what is the meaning of "better"? Do we need both, and if so, in which sequence should they be performed, and how should the results be combined?
- From the literature a variation in applied *threshold values* for classifying LDH is apparent;† should results be corrected for this, or should this merely be made explicit?
- How much was the delay after the first publication on CT in the diagnosis of LDH before an article with sufficient cases and with a mature technique was available?
- How well should MRI perform, in order to be certainly better than CT or myelography in the diagnosis of LDH? What are the constraints for such proof.

Questions on therapy:

- Which are the main complications of CNL and discectomy; is either CNL or surgery safer?
- Is the effectivity of CNL and of surgery sufficiently known, both shortly after treatment and after a few years?
- Is there a role for CNL, and if so is it an alternative for surgery, or an intermediate step between conservative treatment and surgery?
- Is something known about the size of the placebo-effect in LDH treatment?
- Is there a practical way to incorporate the patient's utilities in making choices of treatment?

† The threshold value can be described as the cut-off point on the "continuous spectrum" of *normal radiology* to *distinct LDH on radiological image*, at which a discrimination between *normal* and *pathological* is made. For elaboration, see chapter 3.

References

1 Agre K, Wilson RR, Brim M, McDermott DJ: Chymodiactin post-marketing surveillance — demographic and adverse experience data in 29,075 patients. *Spine* 1984; **9**: 479-485.

2 Am Academy Orthop Surgeons (Chicago): Chymopapain injections win patient approval (News release). Jan 25, 1985.

3 Bell GR, Rothman RH, Booth RE, Cuckler JM, Garfin S, Herkowitz H, Simeone FA, Dolinskas C, Han SS: A study of computer assisted tomography. II. Comparison of metrizamide myelography and computed tomography in the diagnosis of herniated disk and spinal stenosis. *Spine* 1984; **9**: 552-556.

4 Bradford DS, Oegema TR, Cooper KM, Wakano K, Chao EY: Chymopapain, chemonucleolysis and nucleus pulposus degeneration. A biochemical and biomechanical study. *Spine* 1984; **9**: 135-147.

5 Crawshaw C, Frazer AM, Merriam WF, Mulholland RC, Webb JK: A comparison of surgery and chemonucleolysis in the treatment of sciatica. *Spine* 1984; **9**: 195-198.

6 Einarson TR, McGhan WF, Bootman JL, Sabers DL: Meta-analysis: quantitative integration of independent research results. *Am J Hosp Pharm* 1985; **42**: 1957-1964.

7 Ejeskær A, Nachemson A, Herberts P, Lysell E, Andersson G, Irstam L: Surgery versus chemonucleolysis for herniated lumbar discs. A prospective study with random assignment. *Clin Orthop Rel Res* 1983; **174**: 236-242.

8 Fraser RD: Chymopapain for the treatment of intervertebral disc herniation - The final report of a double-blind study. *Spine* 1984; **9**: 815-818.

9 Goldstein G, Gross PF: The treatment of herniated disc in Australia. *Aust Fam Physician* 1985; **14**: 1179-1190.

10 Green DM, Swets JA: *Signal Detection Theory and Psychophysics.* New York, John Wiley & Sons, Inc, 1966. (Reprinted in 1974 by R.E. Krieger Pub Co, Huntington, NY.)

11 Hitselberger WE, Witten RM: Abnormal myelograms in asymptomatic patients. *J Neurosurg* 1968; **28** :204-206.

12 Howe J, Frymoyer JW: The effects of questionnaire design on the determination of end results in lumbar spinal surgery. *Spine* 1985; **10**: 804-805.

13 Javid MJ: Treatment of herniated lumbar disk syndrome with chymopapain. *JAMA* 1980; **243**: 2043-2048.

14 Kassirer JP, Moskowitz AJ, Lau J, Pauker SG: Decision Analysis: A progress report. *Ann Int medicine* 1987; **106**: 275-291.

15 [Lancet; editorial]: Chymopapain and the intervertebral disc. *Lancet* Oct 11, 1986.

16 Lotz PR, Seeger JF, Gabrielsen TO: Prospective comparison of epidural venography and iophendylate myelography in the diagnosis of herniated lumbar disks. *Radiology* 1980; **134**: 127-132.

17 Lusted LB: *Introduction to medical decision making.* Springfield, Ill., CC Thomas, 1968.

18 Martin G: The management of pain following laminectomy for lumbar disc lesion. *Ann R Coll Surg (England)* 1981; **63**: 244-252.
19 Mayer HM, Brock M: Chymopapain-Allergie: die diagnostische Wertigkeit eines Hauttests vor und nach Chemonucleolyse. *Neurochirurgia* 1985; **28**:51-56.
20 Mayfield FH: Complications of laminectomy. *Clinical Neurosurgery* 1976; **23**: 435-439.
21 Mc Rae DL: Asymptomatic intervertebral disc protrusion. *Acta Radiol* 1956; **46**: 9-27.
22 McCulloch JA: Chemonucleolysis: experience with 2000 cases. *Clin Orthop Rel Res* 1980; **146**: 128-135.
23 Pauker SP, Pauker SG: A directive approach to genetic counseling using decision analysis. *Yale J Biol Med* 1977; **50**: 275-289.
24 Pauker SG, Kassirer JP: Clinical applications of decision analysis: a detailed illustration. *Semin nucl med* 1978; **8**: 324-335.
25 Raaf J: Some obeservations regarding 905 patients operated upon for protruded lumbar intervertebral disk. *Am J Surg* 1959; **97**: 388-397.
26 Raiffa H: *Decision analysis.* Reading, MA, Addison-Wesley, 1968.
27 Spangfort EV: The lumbar disc herniation. *Acta Orthop Scand, Suppl. 142, 1972.*
28 Sussman BJ: Inadequacy and hazards of chymopapain injections as treatment for intervertebral disease. *J Neurosurg* 1975; **42**: 389-396.
29 Sutton JC: Chemonucleolysis — current status and future outlook. *Neurochirurgia* 1986; **29**: 173-178.
30 Sutton JC: Chemonucleolysis in the management of lumbar disk disease - a minimum 6-year follow-up evaluation. *Clin Orthop Rel Res* 1986; **206**: 56-60.
31 Teplick JG, Haskin ME: Spontaneous regression of herniated nucleus pulposus. *AJNR* 1985; **6**: 331-335.
32 Tversky A, Kahneman D: Judgement under uncertainty, heuristics and biases. *Science* 1974; **185**: 1124-1131
33 Vlek CAJ, Wagenaar WA: Oordelen en belissen in onzekerheid. P 447-492 in: JA Michon, EGJ Eijkman, LFW de Klerk, *Handboek der Psychomomie.* Deventer, Van Loghum Slaterus 1976.
34 Von Neuman J, Morgenstern O: *Theory of Games and Economic Behavior.* 2nd ed. Priceton, NJ, Princeton University Press, 1947.
35 Watts C: Complications of chemonucleolysis for lumbar disc disease. *Neurosurgery* 1977; **1**: 2-5.
36 Weber H: Lumbar disc herniation. *J Oslo City Hosp* 1978; **28**: 33-64, 89-120.
37 Weinstein J, Spraat KF, Lehmann T, McNeill T, Hejna,W: Lumbar disc herniation. A comparison of the results of chemonucleolysis and open discectomy after ten years. *J Bone Joint Surg* 1986; **68A**: 43-54.
38 Weinstein MC, Fineberg HV, Elstein AS, Frazier HS, Neuhauser D, Neutra RR, McNeil BJ: *Clinical Decision Analysis.* Philadelphia, WB Saunders Co., 1980.
39 White JC: Results in surgical treatment of herniated lumbar intervertebral discs. *Clin Neurosurg* 1966; **13**: 42-54.

1

LUMBAR DISK HERNIATION: DIAGNOSIS WITH CT OR MYELOGRAPHY?

Jaap Schipper, MD[1]
Jan WPF Kardaun, MD [2]
Reinder Braakman, MD, PhD [3]
Krijn J van Dongen, MD, PhD[1]
Gerhard Blaauw, MD, PhD [4]

[1] Department of Diagnostic Radiology,
University Hospital Rotterdam, Dijkzigt

[2] Department of Public Health and Social Medicine,
Erasmus University Rotterdam

[3] Department of Neurosurgery,
University Hospital Rotterdam Dijkzigt

[4] Department of Neurosurgery,
St. Clara Hospital, Rotterdam

The Netherlands.

Acknowledgements: The other members of the Rotterdam Herniated Disc Study include Heleen van Agt, MSc; Marijke van Dishoeck; Geert J Gelpke, MSc; J Dik F Habbema, PhD; Paul J van der Maas, MD, PhD; Ram Singh, MD, PhD; and Mariëtte Westendorp. We thank Dieter N Hüpscher, MD, PhD, for his courtesy in providing the examinations performed in his department and Betty AA Koehorst for preparing the manuscript.
Supported by a grant from the Ministry of Welfare, Health, and Cultural Affairs.

Summary

The value of computed tomography (CT) and myelography as single investigations in the diagnostic evaluation of patients with radiating leg pain probably due to lumbar disk herniation (LDH) has been adequately demonstrated. However, the extent to which CT can replace myelography and the conditions in which the examinations should be combined and in which order are still uncertain. Results of CT scans and myelograms from 461 patients with symptoms of lumbar root compression, probably due to LDH, were evaluated and compared with surgical results, if available. The sensitivity of myelography exceeded that of CT (82% vs. 73%), but its specificity was lower (67% vs. 77%). The positive predictive value of myelography only slightly differed from that of CT (93% vs. 94%). These results were used to establish a sequential diagnostic workup for patients with radiating leg pain. If, in this population with a high prior probability of surgery, CT had been the investigation of first choice in patients suspected of having LDH, the number of myelographic procedures performed could have been reduced by two-thirds.

Index terms: Myelography, 336.112 — Spine, CT, 336.1211 — Spine, diseases, 336.783 — Spine, intervertebral disks, 336.783

Published in: *Radiology* 1987; **165**: 227-231

The two most important investigations now available for detecting lumbar disk herniation (LDH) are myelography and computed tomography (CT) of the lumbar spine. Myelography has been used for decades, and its value in the diagnosis of the cause of sciatica has been adequately demonstrated [1-6]. CT of the lumbar spine has been the subject of many reports during the last few years [7-11]. However, the precise role of CT and its relationship with myelography in the diagnostic workup of patients with a lumbosacral radicular syndrome remains to be defined more clearly [11].

To clarify this relationship, we undertook a study in which data were obtained from a blind evaluation of CT and myelographic results and compared with surgical findings, if available. In this study of a mixed in- and outpatient group of 471 patients with radiating leg pain, the predictive values of both CT and myelography are assessed. A diagnostic scheme is given for patients with radicular radiating leg pain, taking into account both diagnostic accuracy and inconvenience to the patient.

Our study population formed a part of the Rotterdam Herniated Disc Study, which is trying to answer various questions about the efficacy of health care and diagnostic and therapeutic approaches in patients with radiating leg pain possibly due to LDH.

Patients and methods

Patients

Between September 1983 and February 1985, data were collected from all in- and outpatients with radiating leg pain referred to the neurosurgical departments of two Rotterdam hospitals. There were 471 consecutive patients, but records were inadequate in ten. Thus, the study population consisted of 461 patients with a mean age of 43 years, 61% of whom were male. The study included patients with radiating pain in the area of the sciatic or femoral nerve, with or without back pain, with feelings of numbness, or with paresis. A total of 10.5% of the patients had undergone previous back surgery. Patients with a radicular syndrome due to either a tumor of the cauda equina, spine, or pelvis known at entry into the study or due to trauma were excluded. Data from two patients in whom a tumor in the cauda equina was detected during the study were excluded from the analysis. The data collected for every patient included details of complaints and symptoms; physical examination; clinical diagnosis; and findings from plain radiography, CT, or myelography, and, for those who had undergone operations, surgery.

The study design included CT and myelographic examinations that were both to be performed as often as possible in the 461 patients. This could not be done in

all cases, however, because of organizational difficulties in performing CT at short notice and lack of indication for or patient refusal to undergo the invasive myelographic investigation. The results from a few examinations were lost and could not be included in the analysis. The remaining population was categorized into three groups: group 1 consisted of the patients who underwent myelography (n = 339); group 2, those who underwent CT (n = 319); and group 3, those who underwent both CT and myelography (n = 236). Patients in group 3 are also included in groups 1 and 2. The almost equal distribution of therapy and surgical findings among groups 1–3 indicates that the selection of patients to undergo CT, myelography, or both was not clearly influenced by clinical signs and symptoms (Table 1).

After the diagnostic examinations, patients underwent various forms of therapy, including surgery, involving removal of the LDH and contents of the nucleus of the corresponding lumbar disk; chemonucleolysis; or conservative treatment. We used the surgical findings as the standard for deciding about the presence or absence of LDH in our analysis of CT and myelographic findings. At surgery, a lesion was classified as a herniated disk if a nerve root within its sheath was displaced, stretched, compressed, or immobilized by a localized protrusion of disk material. A lesion was also considered to be a herniated disk, even if no compression of a nerve root was evident, in the case of a visualized median or paramedian protrusion of the disk or a defect in the anulus fibrosus.

Simultaneous with this study, a randomized clinical trial took place in one of the two hospitals to compare results of surgery and chemonucleolysis in patients who were candidates for surgery. Some of the patients in our study were included int his trial. In the trial, these patients were divided into two treatment groups, one that underwent surgery and the other, chemonucleolysis. Hence, surgical findings were not available for the latter group of patients. However, this lack of data for the correlation of surgical findings with radiologic findings should not affect the analysis of groups 1–3 because of the almost equal distribution of the type of therapy.

Radiologic examination

All CT examinations were performed with a Philips Tomoscan 350 (Shelton, Conn.) (200 mAs, 120 kV, convolution filter [smoothing factor] 6). Three-millimeter-thick sections were obtained in a plane as parallel as possible to the vertebral end plates. This was achieved by tilting the gantry (maximal inclination 20) and having the patient flex the knees and hips to minimize lumbar lordosis. Lateral scout views were obtained for proper alignment of the sections. Hard copies were made at a window width of 600 HU and a level of 150 HU; if necessary, additional settings were used. The CT examinations were carried out without prior administration of contrast material.

Myelography was performed after exchanging 15 mL of cerebrospinal fluid for 15 mL of Iopamiro 200 (iopamidol; Bracco, Milano, Italy), a water-soluble nonionized iodine contrast material. Images were obtained in at least four directions, supplemented if necessary with tomography, functional views, and a view of the thoracolumbar junction.

Radiologic diagnosis

All CT scans and myelograms were interpreted by the same two physicians in consensus (one was a neuroradiologist experienced with CT and myelography [K.J.v.D.], the other a radiology resident [J.S.]). This evaluation was carried out without knowledge of the clinical signs and symptoms, clinical diagnosis, whether the patient had undergone or would be undergoing surgery, or surgical results. The blind nature of the study was supported by the requirement that at least 3 months had to elapse between the performance and the scoring of CT or myelography.

CT scans and myelograms of the same patient were not reviewed together, and they were evaluated without knowledge of the result of the other examination. Another analysis of the radiologic examinations was carried out with clinical information available; those results will be described elsewhere.

The images were scored with a standard form on which the absence or presence of LDH, bulging disk, and spinal stenosis was noted for each level. The side and site of LDH were also recorded, and stenosis was categorized as developmental stenosis, lateral recess stenosis, or stenosis due to epidural scarring.

The criteria for the presence of LDH on CT scans were (*a*) an asymmetric protruding disk, (*b*) obliteration of the epidural fat, (*c*) compression or displacement of the nerve root, and (*d*) indentation of the dural sac. We used the following myelographic criteria for LDH: (*a*) displacement of the nerve root sheath, (*b*) "amputation" of the nerve root sheath, (*c*) thickening of the nerve root, and (*d*) indentation of the dural sac. A lesion was classified as a bulging disk if the radiologic study demonstrated an extradural sac deformity with symmetric upper and lower margins not extending cephalad or caudad to the level of the intervertebral disk space.

The conclusions of the evaluation were divided into three categories: + = probability of an LDH > 70% (LDH likely); ± = probability of an LDH 30%–70% (LDH doubtful); and — = probability of an LDH < 30% (LDH unlikely).

A radiologic examination result was considered to be correctly predicted when it was confirmed at surgery, independent of the level. We used the following definitions of the different measures of test qualities: (*a*) sensitivity equals the number of correctly predicted LDHs out of all LDHs found at surgery, (*b*) specificity equals the number of correctly predicted negative examinations out of those for all patients without LDH at surgery, (*c*) positive predictive value equals

the number of correctly predicted LDHs out of all positive examinations, and (*d*) negative predictive value equals the number of correctly predicted negative examinations out of all negative examinations. Positive (negative) prior probability is the chance that a patient has (does not have) an LDH before a test (i.e., a radiologic procedure) is applied. It is a measure of how severely the patients are selected before entering the study.

Results

Group 1: Myelography

Of 339 myelograms, 13 were of such quality that they could not be evaluated accurately (e.g., epidural contrast, not enough views obtained). In Table 2 the results of myelography are cross tabulated with the type of treatment and with surgical findings. Surgical confirmation was available in 263 cases; in 229 of these an LDH was found. A myelogram positive for LDH was found in 191 of these 229 cases, resulting in a sensitivity of 83%. In 34 cases, no LDH was found at operation. These 34 cases do not all represent negative explorations; 22 patients had obvious spinal stenosis as a cause of their radicular symptoms and signs. The positive predictive value and negative predictive value were 95% and 39%, respectively.

Table 1 — Treatment and surgical findings in patients with suspected LDH

Treatment	Group 1	2	3
Surgery			
LDH found	229	197	165
No LDH found	34	38	30
Chemonucleolysis	35	29	25
Conservative	28	37	16
Total	326	301	236

See text for definitions of groups 1-3.

Table 2 — Comparison of radiologic scores and test qualities for patients with suspected LDH

A. Radiologic score

Treatment	Group 1 Myelography			Group 2 CT			Group 3 Myelography			Group 3 CT		
	+	±	−	+	±	−	+	±	−	+	±	−
Surgery												
LDH found	191	5	33	140	6	51	136	3	26	120	3	42
No LDH found	10	2	22	8	1	29	10	2	18	7	1	22
Chemonucleolysis	29	1	5	21	0	8	21	1	3	16	3	6
Conservative	8	2	18	10	1	26	6	2	8	2	3	11

B. Test qualities *

	Group 1	Group 2	Group 3 Myelography	Group 3 CT
Sensitivity	191/229 (83)	140/197 (71)	136/165 (82)	120/165 (73)
Specificity	24/34 (71)	30/38 (79)	20/30 (67)	23/30 (77)
Positive predictive value	191/201 (95)	140/148 (95)	136/146 (93)	120/127 (94)
Negative predictive value	24/62 (39)	30/87 (34)	20/49 (41)	23/68 (34)
Positive prior probability	229/263 (87)	197/235 (84)	165/195 (85)	
Negative prior probability	34/263 (13)	38/235 (16)	30/195 (15)	

Numbers in parenthese are percentages. See text for definitions of groups 1-3 and +, ±, and −.
* For calculation of the test qualities, two-by-two tables are calculated by combining negative (−) and equivocal (±) categories and by including only patients who underwent surgery.

Group 2: CT

Of the 319 CT examinations available, 18 were technically inadequate for various reasons (e.g., extreme obesity, inaccurate alignment of the sections). A cross tabulation of results of CT, type of treatment, and surgical findings is presented in Table 2. Surgical confirmation was available in 235 cases, LDH being found in 197 of these. In 140 of these 197 cases, the LDH was predicted by means of CT,

Table 3 — Relationship between CT and myelographyic findings and treatment

Treatment	Both Positive	Negative CT/ Positive Myelography	Positive CT/ Negative Myelography	Both Negative
Surgery				
LDH found	101	35	19	10
No LDH found	1	9	6	14
Chemonucleolysis	14	7	2	2
Conservative	1	5	1	9

resulting in a sensitivity of 71%. Positive predictive value and negative predictive value were 95% and 34%, respectively.

Group 3: CT and myelography

A cross tabulation of the results from both CT and myelography, type of treatment, and surgical findings is shown in Table 2. Surgical confirmation was available in 195 of the 236 cases in which both radiologic examinations were performed. Myelography and CT had sensitivities of 82% and 73%, respectively.

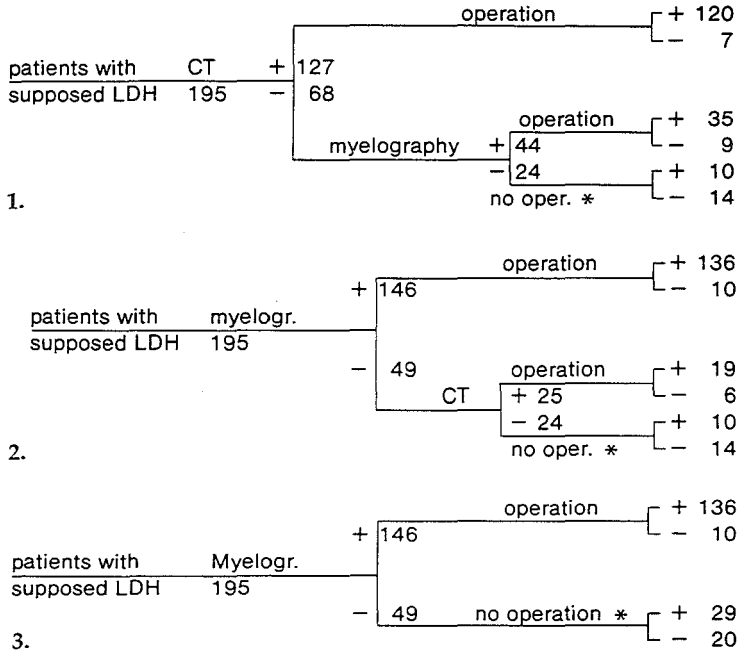
Combined investigations

When the results in group 1 are compared with those in group 3 and when those of group 2 are compared with those of group 3, it is clear that there are only negligible differences. Therefore, the comparison between CT and myelography and the appropriate diagnostic sequence will subsequently be derived from the group who underwent both investigations (group 3).

When equivocal test results are considered as negative, a combined interpretation of CT and myelographic findings has four possible combinations: one unanimously positive, one unanimously negative diagnosis, and two incongruences. Table 3 presents the type of treatment and the surgical findings for these four combinations. The figures in this table are used to analyze the three different diagnostic schemes represented in Figures 1–3.

In pathway 1 (Fig. 1) a CT scan with results positive for LDH immediately leads to surgery, while in the case of a negative CT scan myelography is performed. A positive myelogram would again lead to surgery and a negative myelogram to conservative treatment. In pathway 2 (Fig. 2) myelography is the examination of

Figures 1–3. Diagrams show diagnostic sequences for use of CT and myelography in diagnosis of LDH. CT and myelographic data were available for all 195 patients; these diagrams show what would have happened if the procedures had been applied selectively. Numbers represent number of patients, + = LDH found at surgery, — = LDH not found at surgery, * = these results are not available in practice. (1) Pathway 1. In this pathway, patients undergo surgery when CT findings are positive for LDH. When they are negative, myelography is performed; if myelographic results are positive, surgery is performed. This pathway has a sensitivity of 94% (155 of 165); specificity, 47% (14 of 30); false-positive rate, 8% (16 of 195); and false-negative rate, 5% (10 of 195). (2) Pathway 2. Patients undergo surgery when myelographic findings are positive for LDH. When they are negative, CT is performed, and surgery follows when CT results are positive. Sensitivity, etc., are the same as in pathway 1. (3) Pathway 3 (reference situation). Only myelography is used for diagnosis of LDH. Patients undergo surgery if myelographic findings are positive. This method has a sensitivity of 82% (136 of 165); specificity, 67% (20 of 30); false-positive rate, 5% (10 of 195); and false-negative rate, 15% (29 of 195).



first choice, with CT being reserved for patients in whom myelographic results are negative. In pathway 3 (Fig. 3) a positive myelogram leads to surgery and a negative myelogram to conservative treatment. It must be remembered that these are hypothetical diagnostic sequences: all patients in this group underwent CT, myelography, and surgery.

The numbers in the pathways given in Figures 1–3 represent the number of patients ("reconstructed cases" derived from study data) that would have followed the possible routes and the consequences this would have had for the number of correctly and incorrectly treated patients and the number of required investigations. When CT is the examination of first choice, myelography is performed only in those patients with a negative CT scan (Fig. 1). If in these cases a positive myelogram is considered as an indication for surgery, 94% (155 of 165) of all the present LDHs would be predicted correctly, with a false-positive rate of 8% and a false-negative rate of 5%. This pathway means there would be a reduction of 65% in the number of myelograms obtained. When myelography is the investigation of first choice and CT is the subsequent examination if the myelogram is negative (Fig. 2), the results are the same, but the number of myelograms obtained is approximately three times greater than that obtained in pathway 1. When only myelography is performed (Fig. 3), 82% (136 of 165) of all LDHs will be predicted correctly, with a false-positive rate of 5% and a false-negative rate of 15%.

Discussion

When a patient with radiating leg pain is suspected of having an LDH, radiologic confirmation of this diagnosis is necessary because physical signs and symptoms alone have a poor sensitivity and specificity in locating the exact level of the disorder [12,13]. Once plain radiographs of the lumbar spine have been obtained to detect most of the other pathologic conditions that can cause a radicular syndrome, a decision has to be made between CT and myelography as the investigation of first choice.

Myelography, an invasive procedure, is likely to reveal the presence or absence of nerve root or cauda equina compression in the lumbar spinal canal. It can demonstrate tumors of the cauda equina, and the L-2 to L-3 level is routinely seen. Because it involves lumbar puncture and administration of contrast media, there is a risk of side effects and hospitalization is usually required.

CT is a noninvasive procedure that usually reveals the cause of radicular compression and demonstrates lateral and intraforaminal pathologic conditions well. CT cannot, however, be expected to demonstrate tumors of the cauda equina, and the L-2 to L-3 level is not routinely visualized. There is no need for contrast material administration or hospitalization.

Many authors [14-18] who have compared CT and myelography conclude that CT offers a good alternative to myelography in the diagnostic management of patients with sciatica. As Bell et al. [11] already pointed out, however, almost all these comparative studies had serious design limitations (e.g., absence of surgical confirmation, studies carried out on a biased population).

In contrast to other authors, Bell et al. [11] conclude that myelography should remain the diagnostic procedure of first choice. This difference in conclusion may be explained by the fact that Bell et al. analyzed only patients with a surgically confirmed LDH or spinal stenosis, limiting the evaluation of a negative test outcome and concealing a large part of the false-positive errors. This may well influence the applicability of their interpretation in daily practice, where examinations are also carried out in "healthy" patients. Moreover, when results of CT and myelography are compared with the aim of establishing the priority of one over the other, sensitivity is not the only factor to be considered.

To avoid the bias of analyzing only patients with surgically proved LDH or stenosis and to be able to determine and compare the predictive values of the two examinations, one must collect data from all patients undergoing one or both examinations. To allow the results to be applied as widely as possible to other, similar populations, we attempted to determine the intrinsic value of both CT and myelography. To achieve this, we chose to perform a blinded evaluation of all examinations. In addition, all examinations were interpreted by two radiologists in consensus according to a standard form with criteria for the presence of LDH.

This blinded evaluation avoids to the maximal extent feasible the bias of selecting patients for surgery on the basis of findings from the same procedure. In our analysis, surgical findings were used as the standard for deciding about the presence or absence of LDH. This was done in spite of the well-known problem of observer disagreement during surgery about the existence of and LDH [19]. The use of surgical findings as the standard can be defended because of the subordinate role of radiologic diagnosis to therapeutic (surgical) possibilities. In other words, the ability of a radiologic investigation to demonstrate an abnormality that cannot be found at surgery does not benefit the patient.

A refinement of our analysis would have been to introduce the level-specific agreement between surgical findings and those from radiologic examinations. We did not do this for two reasons. First, when CT scans and myelograms are blindly evaluated, it is sometimes difficult to determine the specific level of the disorder (this is truer for CT than for myelography). However, the problem is considerably diminished in the "clinical" evaluation, with the clinical signs and symptoms and plain radiographs of the patient at hand. Second, the use of global presence or absence as an agreement criterion makes the analysis easier and more definitive and interpretable. We realize, however, that the results would be somewhat poorer if the level-specific criterion were used.

Our present approach is supported by the results of another analysis of our patients (work in progress) in which results from CT and myelography were evaluated with known clinical signs and symptoms but with the requirement of a level-specific agreement. This analysis showed test qualities that are the same or better than those in this study.

Although all examinations were scored for the absence or presence of other causes of radicular symptoms, such as stenosis of the lateral recess of spinal canal and postsurgical epidural scarring, the comparison of CT, myelographic, and surgical findings in our study focuses only on the main diagnosis (LDH) and not on these additional (possibly concurrent) diagnoses. In our population, 22 of all patients who underwent surgery appeared to have a canal stenosis as the single surgical finding. There is a general agreement [11] that CT is equal to or better than myelography for diagnosing spinal stenosis; thus, our conclusions about the sequential use of CT and myelography will not be altered by our focus on LDH.

When CT is compared with myelography, the latter appears to have a clearly higher sensitivity at the expense of a lower specificity. The predictive values of a positive test result are approximately equal for both examinations. The chance that an LDH will be found at surgery when its presence has been predicted with both CT and myelography is almost 100% (101 of 102). Only one patient in our population with both positive CT and myelographic examinations was treated conservatively. In this patient invasive treatment was contraindicated due to a hematologic disorder. Ten patients, each with negative CT and myelographic examinations (when blindly evaluated), proved to have an LDH at surgery.

Although the best results will be obtained when all patients undergo both examinations, it is clear that the two investigations should be combined in a logical sequence to offer a high diagnostic accuracy with the least possible inconvenience to the patient. On the basis of the three different pathways in Figures 1–3, it is clear that when CT is the examination of first choice and myelography is performed only in those patients with a negative CT scan, 94% of all LDHs present will be predicted correctly, with an 8% false-positive rate and a 5% false-negative rate. When only myelography is performed, 82% of all LDHs will be predicted correctly, with a somewhat lower false-positive of 5% and a higher fasle-negative rate of 15%. We conclude from these results that CT of the lumbar spine should be the examination of first choice, with myelography being reserved for those patients in whom the clinical suspicion of the presence of an LDH is not confirmed with CT. In our study population this would have meant a 65% reduction in the number of invasive myelographic procedures, combined with a high true-positive rate for detected pathologic condition.

References

1. Amundsen P, Helsingen P, Kristiansen K. Evaluation of lumbar radicolography ("myelography") with water soluble contrast media. *Acta Radiol [Diagn] (Stockh)* 1963; 1: 659-665.
2. Cook PL, Wise K. Correlation of the surgical and radiographic findings in lumbar disc herniation. *Clin Radiol* 1979; **30**: 671-682.
3. Grainger RG, Kendall BE, Wylie IG. Lumbar myelography with metrizamide: A new non-ionic contrast medium. *Br J Radiol* 1976; **49**: 996-1003.
4. Hirsch C, Rosencrantz M, Wickbom I. Lumbar myelography with water soluble contrast media. *Acta Radiol [Diagn] (Stockh)* 1969; **8**: 54-64.
5. Hudgins WR. Predictive role of myelography in the diagnosis of ruptured lumbar discs. *J Neurosurg* 1970; **32**: 152-162.
6. Simon W. Stimmen Röntgendiagnose und Operationsbefund bei der lumbalen myelographie überein? *Radiologe* 1981; **21**: 347-352.
7. Carrera GF, Williams AL, Haughton VM. Computed tomography in sciatica. *Radiology* 1980; **137**: 433-437.
8. Lackner K, Schroeder S. Computertomographie der Lendenwirbelsäule. *RöFO* 1980; **133**: 124-131.
9. Novetsky GJ, Berlin L, Epstein AJ, Lobo N, Miller SH. Extraforaminal herniated disc: Detection by computed tomography. *AJNR* 1982; **3**: 653-655.
10. Williams AL, Haughton VM, Syvertsen A. Computed tomography in the diagnosis of herniated nucleus pulposus. *Radiology* 1981; **135**: 95-99.
11. Bell GR, Rothman RH, Booth RE, et al. Comparison of metrizamide myelography and computed tomography in the diagnosis of herniated lumbar disc and spinal stenosis. *Spine* 1984; **9**: 552-556.
12. Edgar MA, Park WM. Induced pain patterns on passive straight leg raising in lower lumbar disc protrusion. *J Bone Joint Surg [Br]* 1974; **56**: 658-667.
13. Hakelius A. Prognosis in sciatica. *Acta Orthop Scand* 1970; **S29**: 1-76.
14. Anand AK, Lee BCP. Plain and metrizamide CT of lumbar disc disease: A comparison with myelography. *AJNR* 1982; **3**: 567-571.
15. Haughton V, Elderik O, Magnaes B, Amundsen P. Prospective comparison of computed tomography and myelography in the diagnosis of herniated lumbar disc. *Radiology* 1982; **142**: 103-110.
16. Moufarrij NA, Hardy FW, Weinstein MA. Computed tomographic, myelographic, and operative findings in patients with suspected herniated lumbar disc. *Neurosurgery* 1983; **12**: 184-188.
17. Raskin SP, Keating JW. Recognition of lumbar disc disease: Comparison of myelography and computed tomography. *AJR* 1982; **139**: 349-355.
18. Sachsenheimer , Hamer J, Müller HA. Value of spinal computed tomography in diagnosis of herniated lumbar discs. *Acta Neurochir (Wien)* 1982; **60**: 107-114.
19. Hirsch C, Nachemson A. Reliability of lumbar disc surgery. *Clin Orthop* 1963; **29**: 189-195.

2

CT, MYELOGRAPHY AND PHLEBOGRAPHY IN THE DETECTION OF LUMBAR DISK HERNIATION

An analysis of the literature

Jan WPF Kardaun [1]
Jaap Schipper [2]
Reinder Braakman [3]

[1] Department of Public Health and Social Medicine,
Erasmus University Rotterdam

[2] Department of Radiology,
Leyden University Hospital, Leyden

[3] Department of Neurosurgery,
University Hospital Dijkzigt, Rotterdam

The Netherlands

Summary

Despite the large number of reports on the relative usefulness of various radiographic procedures for the diagnosis of lumbar disk herniation, there has been no consensus of opinion on the best imaging procedure. Different study designs, including criteria for patient selection and retrospective consideration of patients who underwent surgery only, hamper direct comparisons between studies. A major drawback is the common use of "accuracy" as a measure of quality. We reviewed the CT, myelographic, and phlebographic findings in lumbar disk herniation published since 1970. After the reports were systematically classified and assessed for quality, the results became more coherent.

Many results tend to be sensitive and not very specific. We found that there was no clear difference in the overall diagnostic quality of phlebography, myelography, and CT.

Published in the American Journal of Neuroradiology 1989; **10**: 1111-1122

Introduction

The most common causes of radicular compression syndrome are a herniated intervertebral disk and stenosis of the spinal canal. Over the years the ability of several radiologic investigations like myelography, epidural phlebography, CT, and more recently MR imaging to detect either of these abnormalities has been studied. However, there has been no consensus of opinion on the best imaging procedure. Differences in study design and methods of analysis have hindered comparisons.

The objective of this article was to assess the extent to which the capability of CT, myelography and phlebography to visualize lumbar disk herniation can be derived from the literature and to explain the discrepancies in results. To this aim, special attention was paid to the way the results in the different studies were obtained.

Materials and methods

Selection of literature

The advent of safe water-soluble contrast media limits the literature review to 1970 and later. The first CT articles appeared in 1976 [1] and 1977 [2]. CT developments have been very rapid, especially since 1983.

There is another X-ray based procedure for diagnosing herniated disks, phlebography, that despite its history of about 30 years [3], never became as popular as myelography. When catheterized [4,5], selective [6,7], and later double-sided catheterized [8] phlebography were introduced, the results were claimed to be equal or superior to myelography. However, since 1980, CT has a more dominant place in the literature than phlebography. Though it seems that phlebography has been overtaken by the newer technologies, it is included in this review to determine its relative diagnostic accuracy. Because MRI is still evolving, it is not included in this review, unless it is included in an article that is cited for other reasons.

A minimum number of 20 patients in a study was required for inclusion of an article in the analysis, unless the article announced an innovation. Only publications in regular journals were considered — no proceedings, textbooks, or monographs were searched.

There is an important difference between assessing the accuracy of one department or one study and trying to assimilate the results from several publications into a more general statement. Therefore, before presenting the concrete results of our review, we will discuss the requirements of a study if it is to play a useful part in comparative literature.

Requirements of the literature

The requirements of the literature of the reviewed investigations can be considered from different viewpoints. Several aspects are evaluated: (1) the selection criteria of the patient; (2) the pro- or retrospective design; (3) the evaluation criterion of the test, that is, the verification of the diagnosis; and (4) the measure for the results. These points are discussed in reverse order, as the reasons for the first steps in the design of a study are best understood by knowing the aim of the investigation.

The measurement of the predictive power of diagnostic procedures. — In the reviewed literature, the usual measure was "accuracy", that is, the number of all correct diagnoses divided by the total number of diagnoses. A correct diagnosis can be a suspected herniation at myelography that is also found at operation. A correct diagnosis can also be the assumption of absence of herniation on a radiologic procedure, which is confirmed either by operation (e.g. on clinical grounds or because of a positive CT) or by the transitory and noniterative character of the complaints. Accuracy is a sufficient measure to determine whether a particular physician or department makes many or few errors, but lumping together these two types of correct results hides important and interesting information: We do not know whether the accuracy pertains to a population with many or only few diseased persons.

Assume, for example, that a population has a 90% prevalence of herniated disk (as is realistic after screening by clinical signs and other means, e.g. complaints that are resistant to a "lege artis" conservative therapy), then it would be a good guess when a myelogram was completely uninformative (e.g., was lost) to pretend that a herniated disk was seen. In fact, an accuracy of 90% could be achieved (and a lot of time saved) by completing the radiographic reports for all patients as having a positive myelogram, before the myelogram was seen. (This is only a hypothetical example, of course.)

	T+	T−	
D+	*a*	*b*	*p*
D−	*c*	*d*	*q*
	m	*n*	*N*

We need to introduce a few symbols and the concept of the two-by-two table to determine what would be a better measure of the predictive power of a procedure. Assuming that a disease (herniation, stenosis) is either present or absent (D+ and D–) and that a diagnostic procedure did or did not predict the same disease (T+

and T– for a test, i.e., myelography, CT, or phlebography), there are four possible combinations for each test.
In a clinical context *a, b, c,* and *d* are usually called the true-positive, false-negative, false-positive, and true-negative test results. The accuracy of the procedure is usually defined as $(a+d)/N$, the prevalence of the disease is p/N, the specificity is d/q, the sensitivity a/p.

We can now easily see why the accuracy is only a limited measure: it does not tell us anything about sensitivity and specificity. It makes all the difference whether the erroneous myelography reports are concentrated in *b* or in *c*, for in the first case we have lower sensitivity and in the other a lower specificity. The quadruple specificity, sensitivity, prevalence, and *N* completely describe the two-by-two table.

There are more ways to look at this table: usually a practicing clinician is confronted with the question: If my test is positive (or negative), what is the probability of disease or no disease? The probabilities a/m and d/n are usually referred to as the predictive value of a positive test ($PV+$) or a negative test ($PV-$), respectively. In order to describe the table completely, we need in addition to the $PV+$ and $PV-$ (and N) a fourth quantity, such as the ratio m/n. We are more interested in the prevalence (p/N) of a disease than in this m/n because it has a more direct interpretation.

For all these ways of assessing the effectiveness of a radiographic investigation, all four cells in the two-by-two table must be known. This not only follows the above simple calculations, but it also common sense. If we want to know the conclusions that can be drawn from a "positive" or "negative" radiograph, then both categories of patients should be followed to see what is really the (likely) cause of the radicular syndrome.

This leads immediately to two other requirements: the evaluation of the patients and the pro-/retrospective study design.

Evaluation of the patient. — All patients who had myelography or CT should have verification of their diagnosis. For those patients likely to have herniation, surgical verification often will be available, but for others at least an attempt should be made to see whether their follow-up gives an indication of the presence or absence of herniation. Of course, this is only an approximation of the anatomic situation, but it is better than ignoring the issue at all, and it often will give a fair clue to the right answer.

Pro- or retrospective study design. — The above leads naturally to the requirement that studies for evaluating diagnostic test procedures ought to be "prospective", that is, they should assess all patients who present themselves to the study. (Retrospective studies are to be used to determine etiology when we want to know whether some event/exposure caused a disease.) The question is not what to do with patients in whom herniation was found, but to know what to do with a patient with signs and history that are likely to be due to a herniation.

Patient selection. — There is a wide range of indications for studying patients with CT, myelography, or phlebography because of radicular compression complaints. In some centers, all patients with suspected herniation will have one of these investigations. In other centers, only patients with clear symptoms who are to undergo surgery routinely have a CT scan or myelogram. In yet other centers, patients with very clear symptoms undergo surgery on clinical presentation only. To compare information about history and clinical signs in reports from several centers, it is useful to classify each patient into one of a few groups varying from slight to clear clinical evidence of herniation.

Blind vs clinical information. — It is important to know whether the radiographic reports are completed with or without the use of clinical information and other tests. Depending on the goal of the analysis, both ways can be defended. The combined use is good if one wants to know whether this combination is sufficient for making the decision to operate (and where) or not to operate, and if one is interested in an analysis that should be more applicable in a clinical setting. The blinded approach is preferable if one wants to know the contribution of each test in the decision making.

Classification of the articles

The articles selected for review were classified with respect to the following points (Table 1): (1) whether they included, in some way, the use of clinical, myelographic, CT, phlebographic, and surgical data; (2) what selection criteria were used for inclusion of patients in the study; (3) whether there was an apparent bias in the selection; (4) whether the verification criteria for the diagnosis were strict or lax; (5) whether the data were, either in the text or in a table, presented in a complete way, so that a complete two-by-two table could be (re)constructed; (6) if a table could be reconstructed, the prevalence, sensitivity, and specificity of the test; (7) year of publication; (8) number of patients in the study (if some group of patients was not analyzed at all, but excluded right away, then the number in the table may be lower than that mentioned in the article or title); (9) whether the test was evaluated blindly or together with clinical information; (10) whether the analyses were done per patient, or per level; and (11) how equivocal tests were handled. Many articles analyzed several groups of patients, in most cases the smaller group was a subgroup of the larger.

Results

Comments on selected reports

The relevant literature is summarized briefly and evaluated relative to our research question. Though almost all reports did provide at least some measure of prediction, mostly accuracy, that often was not the main point of interest of the articles. In articles with an emphasis on technical, anatomic, or differential diagnostic aspects of a procedure, or on complications, rare cases, etc., the presentation of the numeric data and the accuracy of the radiologic procedures was often secondary. By looking mainly at the presentation of the data and reproducibility of the predictive power of the diagnostic method, it is not possible to give full credit to the overall quality and importance of the articles. The articles are summarized chronologically, highlighting new information not included in earlier papers.

In 1970, Hudgins [9] analyzed 490 patients admitted for low back pain or leg pain and 102 patients who had lumbar myelograms for other reasons (controls). Hudgins asserted that most articles on this subject use the wrong approach, that is, they study a population of (surgically) proved herniations and do not assess the predictive value of a radiographic procedure. Besides having controls, Hudgins followed all operated patients. In addition to giving the predictive value of myelography and giving complete data, several interesting questions were discussed that are outside the scope of this review.

In 1974, Gargano et al. [6] analyzed 32 patients who had both myelograms and phlebograms and who underwent surgery for herniated disks. The emphasis of this article was on anatomy of and indications for phlebography, and much of the older literature was reviewed. A tabulation of the data showed better results for phlebography, but their population consisted of patients who previously had "clinical lumbar disk disease, but negative or equivocal myelograms". This gave a bias in favor of phlebography.

In 1976, MacNab et al. [11] studied a group of 110 patients with symptoms of herniation, who underwent phlebography. They considered the myelograms of the 50 patients without prior disk surgery and found a diagnostic accuracy of 98% and 90% for phlebography and myelography, respectively. Their emphasis was more on anatomy and phlebographic technique than on numeric evaluation. Their data could not be reconstructed.

In 1977, Mohsenipour et al. [13] reviewed the myelograms of 500 patients. Three hundred seventy-four patients with a clearly positive myelogram underwent surgery, as did 27 of 85 patients with an equivocal myelogram. It is unclear whether none of the 41 with a clearly negative myelogram underwent surgery on clinical grounds. A few complications were reported, and a low threshold was favored for obtaining a myelogram, arguing that clinical signs cannot predict the level of a

TABLE 1: Summary of Literature Assessing Imaging Methods in Lumbar Disk Herniation

Year of Publication Reference	Sources Considered	No. and Units of Observations	Direct Comparisons	Blinded Evaluations	Apparent Favoring Bias
1970					
(1) Hudgins [9]	M,CI,S,FU	309 pts	M/S+FU	Not known	NA
		135 pts	M/S	Not known	NA
1974					
(2) Gargano et al. [6]	M,P,S	32 pts	M/S	Not known	P
		32 pts[d]	P/S	Not known	P
1976					
(3) Drasin [10]	P,S	19 levels	P/S	Not known	NA
(4) MacNab et al. [11]	M,P,S	50 pts	M/S	Not known	None apparent
		50 pts[d]	P/S	Yes	None apparent
(5) Miller [12]	P,S	38 pts	P/S	Not known	NA
1977					
(6) Mohsenipour et al. [13]	M,S	401 pts	M/S	Yes	None apparent
(7) Moringlane et al. [14]	M,S	140 pts	M/S	Not known	NA
1978					
(8) Roland et al. [15]	M,P,S	111 pts	M/S	Not known	P
		111 pts[d]	P/S	Not known	P
1979					
(9) Cook and Wise [16]	M,S,PR	62 levels	M/S	Not known	NA
(10) Meyenhorst [8]	M,P,S[j]	339 levels	P/M	Not known	None apparent
		151 levels	M/S	Not known	None apparent
		151 levels[d]	P/S	Not known	None apparent
1980					
(11) Lotz et al. [17]	M,P,S	37 levels	M/S	No	None
		37 levels[d]	P/S	No	None
		37 levels[d]	M/S	Yes	None
		37 levels[d]	P/S	Yes	None
		37 levels[d]	M/S	No	None
		37 levels[d]	P/S	No	None
(12) Thijssen et al. [18]	M,S	104 pts	M/S	No	NA
1981					
(13) Gulati et al. [19]	PCT,MCT,M	15 levels	PCT/M	Not known	M
		15 levels[d]	CTM/M	Not known	M
1982					
(14) Anand and Lee [20]	PCT,MCT,M	25 pts	PCT/M	Yes	None apparent
		75 pts	CTM/M	Yes	None apparent
(15) Claussen et al. [21]	M,CT,S	41 pts	CT/M	Not known	None apparent
		26 pts	CT/S	Not known	None apparent
		23 pts[d]	M/S	Not known	None apparent
(16) Fries et al. [22]	M,CT,S	192 levels	M/S	Not known	None apparent
		227 levels	P/S	Not known	None apparent
(17) Haughton et al. [23]	CT,M,S	55 levels	M/S	Not known	None
		55 levels[d]	CT/S	Yes	None
(18) Jepson et al. [24]	M,S,CI[n]	55 pts	M/S	Not known	NA
		55 pts[d]	M/S	No	NA
(19) Sachsenheimer [25]	M,CT	35 pts	M/CT[o]	Not known	None apparent
1983					
(20) Moufarrij et al. [26]	M,CT,FU	50 pts	CT/S[p]	Not known	None apparent
		46 pts[d]	M/S[p]	Not known	None apparent

Selection Criteria	Treatment of Equivocal Results	Severity of Diagnostic Criteria	Prevalence	Sensitivity	Specificity
Combination[a]	Not stated	Medium	–	–	–
Operated pts	Not stated	Medium	0.79[b]	0.75	0.90
Clinical findings[c]	Not stated	Medium	0.94	0.63	0.50
Clinical findings[c]	Not stated	Medium	0.94	0.93	0.50
Operated pts[e]	Own category	Medium	0.84	0.94	0.33
Operated pts[f]	Not stated	Medium	0.74	0.95	0.77
Operated pts[f]	Not stated	Medium	0.74	0.97	0.85
Operated pts[e]	Not stated	Medium	0.92	0.86	1.00
Operated pts[g]	Not stated	Medium	0.93	0.95	0.36
S candidates[h]	Not stated	Severe	0.96	0.98	0.20
Unclear	Not stated	Medium	0.91[i]	0.40	0.40
Unclear	Not stated	Medium	0.91[i]	0.93	0.60
S verified dx	Own category	Severe	0.84	0.96	0.70
Unclear	Own category	Several	–	–	–
Operated pts	Own category	Several	0.52[k]	0.81	0.94
Operated pts	Own category	Several	0.52[k]	0.97	0.87
Operated pts	Own category	Medium	0.86[l]	0.81	1.00
Operated pts	Own category	Medium	0.89[l]	0.94	0.50
Operated pts	Own category	Medium	0.85[l]	0.71	0.91
Operated pts	Own category	Medium	0.84[l]	0.71	0.42
Operated pts	Own category	Medium	0.85[l]	0.75	0.82
Operated pts	Own category	Medium	0.84[l]	0.74	0.67
Operated pts[m]	Own category	Medium	0.91[b]	0.98	1.00
Cl + M signs	Own category	Medium	–	–	–
Cl + M signs	Own category	Medium	–	–	–
Unclear	Not stated	Medium	–	–	–
Unclear	Not stated	Medium	–	–	–
Susp LDH	Not stated	Medium	–	–	–
Operated pts	Not stated	Medium	0.92	0.88	0.50
Operated pts	Not stated	Medium	0.91	0.81	0.50
S verified dx	Not stated	Medium	0.91[l]	0.87	0.89
S verified dx	Not stated	Medium	0.90[l]	0.92	0.78
Back pain/sciatica	Own category	Medium	0.55	0.93	0.64
Back pain/sciatica	Own category	Medium	0.55	0.97	0.68
Operated pts	None	Medium	0.89	0.90	0.83
Operated pts	Not stated	Severe	0.89	0.61	0.83
S verified dx	Own category	Medium	–	–	–
Operated pts[q]	Not needed[r]	Medium	0.90	0.62	0.80
Operated pts[q]	Not needed[r]	Medium	0.85	0.82	0.43

TABLE 1: (continued)

Year of Publication Reference	Sources Considered	No. and Units of Observations	Direct Comparisons	Blinded Evaluations	Apparent Favoring Bias
1984					
(21) Bell et al. [27]	M,CT,S	122 pts	M/S,CT/S[e]	Yes	None apparent
(22) Bosacco et al. [28]	Cl,CT,M,S	134 pts	Cl/CT/M/S	No	None apparent
		52 pts[d]	CT/S	No	None apparent
		52 pts[d]	M/S	No	None apparent
(23) Valat et al. [29]	P,S	104 pts	P/S	Not known	NA
1985					
(24) Kampmann et al. [30]	M,CT,S	36 levels[v]	CT/M	Not known	None apparent
		31 levels	M/S	Not known	None apparent
		122 levels[v]	CT/S	Not known	None apparent
1986					
(25) Modic et al. [31]	M,CT,MR	151 levels	CT/MR	Yes	None apparent
		218 levels	M/MR	Yes	None apparent
		42 pts	CT/S	Yes	None apparent
		45 pts	M/S	Yes	None apparent
		62 pts	MR/S	Yes	None apparent

Note. — M = myelography; Cl = clinical; S = surgery or surgically; FU = follow-up; P = phlebography; PR = plain radiography; PCT = plain CT; MCT = metrizamide CT; MR = MR imaging; pts = patients; NA = not applicable (only one comparison in the study); dx = diagnosis; Susp LDH = suspected lumbar disk herniation.

a Combination of controls (e.g., patients having myelography for other reasons) and surgically and conservatively treated sciatica patients.

b Data reconstructed by reviewers.

c Patients with clinically suspected herniated disk, but "negative or equivocal" myelograms.

d Same as or included in patient or level group immediately above.

e With normal or equivocal myelograms.

f Operated patients with myelograms and phlebograms, but without prior surgery.

g Includes only patients with positive or dubious myelograms; it is not stated whether patients with negative myelograms had surgery.

h Most patients had had unsuccessful conservative treatment.

i Data also specified by level.

j Distinction made between single- and double-sided phlebography; the data in the table reflect double sided phlebography.

k Data reconstructed by reviewers. Dubious and nondiagnostic cases split up by marginals.

Selection Criteria	Treatment of Equivocal Results	Severity of Diagnostic Criteria	Prevalence	Sensitivity	Specificity
S verified dx	Weighted	Mixed[t]	–	–	–
Susp LDH[q]	Not stated	Medium	–	–	–
Susp LDH[q]	Not stated	Medium	0.96	0.92	1.00[u]
Susp LDH[q]	Not stated	Medium	0.96	1.00	0.50[u]
Operated pts	Not stated	Medium	0.95[l]	0.99	0.20
Sciatica[n]	Not stated	Medium	–	–	–
Operated pts	Not stated	Medium	0.90	0.96	0.33
Operated pts	Not stated	Medium	0.91	0.99	0.57
Susp LDH[w]	Not stated	–	–	–	–
Susp LDH[w]	Not stated	–	–	–	–
Operated pts[w]	Not stated	Mixed[x]	1.00[y]	0.83	–
Operated pts[w]	Not stated	Mixed[x]	1.00[y]	0.72	–
Operated pts[w]	Not stated	Mixed[x]	1.00[y]	0.83	–

l Detailed data (specified by level and observer) were aggregated by the reviewers for all variants of Lotz et al. Equivocal results were divided, according to marginals, into positive and negative diagnosis. Comparisons 1 and 2 = original clinical presentation; comparisons 3 and 4 = blind interpretation; comparisons 5 and 6 = reinterpretation with clinical information.

m "This paper presents ... the findings of a series of lumbar myelographies of adequate quality".

n Data also available for patients who had surgery on clinical grounds only. This group was quite similar to that with myelography; those patients are not included here.

o No distinction between true positive and true negative.

p Extracted from four-way table.

q Patients with previous surgery or spinal stenosis were excluded.

r An equivocal category was not very important since several diagnoses were considered.

s Data were not tabulated because there was no distinction between true positive and true negative, nor between false positive and false negative.

t Severe, medium, and relaxed.

u Based on two cases.

v There were no errors in CT levels.

w Patients with prior surgery or known diagnoses were excluded.

x Severe and medium.

y There were no negative explorations, making the specificity unknown.

herniation well. The interpretation of a partial or complete block was elaborated with instructive images. Their data are not really complete, but a reconstruction of their accuracy is 72%. From the context, it is assumed that they applied a strict criterion as to the level of herniation.

Also in 1977, Moringlane et al. [14] reviewed 140 patients who underwent operations for herniated disks. They described the population by specifying that all of the patients had had (several) conservative treatments, except for the patients with emergency symptoms. The patients agreed before myelography to undergo surgery if an abnormality were to be found. This was probably a population with a high prevalence. Complete data were presented, and the accuracy was 95% (using strict criteria as to level, side and cause). The authors mentioned the use of the prior probabilities for discriminating between different levels and mentioned for the first time in a numerical analysis the findings at myelography and/or operation of multiple herniations (46% of their patients). They described the myelographic and surgical techniques and provided useful illustrations and differential diagnoses of the myelographic findings.

In 1978, Roland et al. [15] described phlebographic anatomy, technique, and diagnostic criteria. For 111 patients (chosen from 240 with unstated indications for selection) with "mainly ... previous ambiguous or normal myelography", they compared the diagnostic performance of phlebography and myelography. Both per patient and per level analysis were provided, and all the data were presented. This comparison was heavily biased against myelography, but the authors correctly concluded that additional phlebography is useful for the five types of cases they mention.

In 1979, Cook and Wise [16] compared the findings of plain radiography and myelography in 50 operated patients, 49 of whom had lumbar disk protrusions. It is unclear how the patients were selected. Some patients had more than one herniation. There were some negative explorations, but it is not clear, whether this was caused by multiple-level laminectomies. Of 36 patients who had myelography, but no surgery, complications were reported, and related to the use of oil- and water-based media. No follow-up was undertaken to find additional signs of herniation. There was a detailed discussion of the discrepancies in their material, and the literature was reviewed, with some emphasis on technique. They elaborated on how to interpret discrepant findings in terms of clinical usefulness and pointed out that equivocal findings should be considered incorrect as far as clinical use is concerned. With these strict criteria they had an accuracy of 92% (based on levels).

Also in 1979, Gershater and St. Louis [32] analyzed 1200 patients suspected of having lumbar disk herniations who underwent phlebography. The emphasis was on anatomy, technique, differential diagnosis, and complications of phlebography. It is unfortunate that the data in this very large series were insufficient to allow

calculations of the test qualities, for example, the results of 243 phlebograms of nonoperated patients, the selection criteria for surgery (only 50% of all patients were operated), and the correctly negative findings.

Another 1979 publication, by Meyenhorst [8], gave a detailed analysis of anatomy and methods for phlebography, and provided extensive results of 63 surgically verified cases studied with both myelography and phlebography. In 113 patients, the relationship between the myelographic and phlebographic findings was assessed. There was a good analysis of which questions are important for the evaluation of the diagnostic value of radiographic procedures and a good literature review. Tabular material was ample, but it was hard to synthesize the many single aspects of the comparisons. The accuracies were 94% and 84% for double-sided phlebography and myelography, respectively.

In 1980, Lotz et al. [17] argued that many previous reports comparing myelography and phlebography concerned biased populations and did not provide surgical findings. They obtained examination results from 50 patients with clinical signs of herniated disks at L_4-L_5 or L_5-S_1 and evaluated the results with and without clinical information. There was no follow-up of non-operated patients. The data presented were concise and complete. Not counting unsuccessful and equivocal radiographic findings, accuracies ranged from 69 to 89% for myelography and 59 to 89% for phlebography. Considering complications and economic factors, they considered myelography to be the procedure of first choice.

Also in 1980, Thijssen et al. [18] analyzed the myelograms of adequate quality of 104 patients who underwent surgery and partially analyzed the data of 143 patients who did not undergo surgery. It was not stated how many inadequate myelograms were discarded. These authors argued that myelography can be equal to phlebography, provided the technique is of high quality. Attention was given to the inconclusive results of myelography and phlebography, both in their findings and in the literature. An accuracy of 93-98% was stated, depending on the treatment of the inconclusive group. A complete tabulation of the data was provided.

Another 1980 article, by Williams et al. [33], analyzed 16 patients studied with CT who had herniated disks at operation, and mentioned 21 patients who had CT but did not undergo surgery. Since they had no false positives, nor "did false negatives come to their attention", their analysis was oriented towards technical details, exact localization of the herniation, differential diagnosis, and advantages or disadvantages of CT vs myelography.

In 1981, Gulati et al. [19] described 10 patients with clinically herniated disks and at least one positive level at metrizamide myelography. All these patients had a high-resolution CT study. This is one of the first articles on CT for lumbar herniated disk that provided a numeric analysis. The findings were given for eight operated patients; the other two did not have follow-up. All operated patients had

positive CT findings (with perhaps one exception which they classified as an artifact).

Also in 1981, a short, but clear article by Hanson [34] described 22 patients with clinically suspected disks who underwent CT. Although only 22 patients were studied, the study was complete in that the findings of both the seven operated patients and the follow-up of the 15 nonoperated patient were provided. The merits of CT over myelography were discussed.

Another 1981 article, by Hanley et al. [35], reviewed 81 patients who had CT because of suspected disk herniation and did not have prior disk surgery. Twenty-six patients underwent surgery, 16 of whom had positive and 10 of whom had negative CT findings. No follow-up was provided for the others. These authors stressed the importance of fragmented disks and made a plea for the distinction between normal, bulging and herniated disks or equivocal results for the procedures. Their data could not be reconstructed or recalculated.

In 1982, Anand and Lee [20] analyzed 100 patients with suspected lumbar disk disease. All of the patients underwent myelography and plain or metrizamide CT. In their comparison of low-dose metrizamide CT and plain CT, the former was found to be superior. Criteria for the CT diagnosis of herniated disk were refined. All the radiographs were blindly reviewed by two observers and complete data were provided on the comparison of CT with myelography (75% agreement), but the analysis of the 53 operated patients was not very clear. There was no follow-up of the other 47 patients. Because the surgical findings were all positive, it is hard to assess the false-positive rate of the procedures. When they stated that "in our study, when CT and myelography agreed ... there was no problem in diagnosis, and either test could have been used", this cannot be applied to three patients in both the table and text who had negative CT, negative myelography and positive operations.

Also in 1982, Claussen et al. [21] presented the CT data of 77 patients, of whom 41 had myelograms and 26 underwent surgery. Twenty-three of the operated patients had two investigations, for which complete data are presented. There was a general introduction on the research question, and attention was given to CT anatomy and criteria.

In another 1982 article, Dublin et al. [36] reviewed the results of plain film metrizamide myelography, CT metrizamide myelography and plain CT. They found 106 patients with both plain film and CT metrizamide myelograms out of 736 with spinal CT scans. In about half the 106 patients, thoracic or cervical levels were studied. The plain film CT and metrizamide myelograms were evaluated blindly and compared with each other and with clinical/surgical findings. The authors found that CT metrizamide myelography was superior to plain film metrizamide myelography in 40%, equal in 50%, and inferior in 10% on the basis of a decision-oriented criterion, that is, the findings should change the surgical

approach or the clinical management of the patient. They stressed the importance of a high-resolution unit and provided several reasons for problems of interpretation. One of the most interesting remarks was that with CT the patient is investigated in the position most likely to relieve pain. The data were not included in Table 1 because their problem was so different from that in the other studies that their classifications were not comparable, and their selection of patients for the investigation was not clear. Though they provided a schema for the management of spinal neurologic problems, they did not indicate whether CT or plain film metrizamide myelography should be performed first and in what conditions the other investigation should be performed. At least for the plain CT vs CT metrizamide myelography analysis, it was suggested that CT metrizamide myelography be performed only after nondiagnostic plain CT. Their high proportion of osseous hypertrophy and neoplasms suggests a different population from that in the other studies.

A fourth 1982 article by Eldevik et al. [37], compared the reports of CT and myelography with and without clinical information. For 107 previously reported patients (52 were operated), they completed their data with clinical and blind reports for CT and myelography, respectively. There was special emphasis on the difficult diagnostic evaluation of patients with prior surgery. Both CT and myelography were interpreted more accurately without the clinical information. For all 107 patients the differences between the blind and clinically reported studies were compared. Blind studies were better in detecting herniation (for CT, narginally significant; for myelography not significant). For the 52 operated patients, the operative findings were compared, and it was found that both CT and myelography had better results without clinical information as CT had three more false-positive findings with clinical information and myelography had two more false-positive findings; however, they did not account for the fact that myelography with clinical information played a role in the indication for operation so that false-negative findings resulting from clinical information were less likely to be noted. This was not true for CT, but in the case analysis they accounted for five of the six differences, allowing us to classify those as three more false positives and two more false positives, which is not impressive. Their data did not allow for a complete reconstruction.

In 1982, Fries et al. [22] selected 188 patients (244 intervertebral spaces) with both CT-diagnosed and surgically confirmed herniated disks. The false-negative and false-positive findings consequently were limited to multiple-level patients. However, there was a detailed analysis of the distribution of herniation sites over levels and in levels. The capability of CT to diagnose central and migrated herniations was assessed. Techniques, interpretation of images, and an explanation for erroneous CT and myelograms were elaborated. Complete data were provided for CT and myelograms vs operations, and the problems of multiple locations and

multiple diseases were addressed, but their claim of a complete lumbar spine examination (above L_4-L_5) is not supported by the use of clinical indications for the levels.

In another 1982 article, by Haughton et al. [23], data were collected on 107 patients referred for low back pain or sciatic pain. CT and myelographic findings were available for all patients, and operative findings were available for 52. Complete data were provided about the correlation and discrepancies of the diagnosis. Their population did not include patients with normal findings at surgery. The CT findings were evaluated blindly, and myelography was evaluated with clinical information. Strict criteria (site, level, and pathology) were applied, but good accuracies were achieved (84% and 88% for CT and myelography, respectively), which can be related to the probably strict criteria for operation. They argued that the most common differential diagnostic problem is formed by herniation vs bulging anulus and spondylolysis, but argued that perhaps a nuclear fragment lodged behind the posterior ligament can be mistaken for a bulging anulus at surgery.

Also in 1982, Herkowitz et al. [38] compared myelography and phlebography in 30 patients with a surgically verified diagnosis of herniated disk or spinal stenosis. They clearly defined what constitutes herniation and stenosis and stated the correct (i.e. usual) definitions of sensitivity and specificity, but their tables used a less obvious interpretation of these rates, so their figures cannot be compared. Fortunately, they provided a table presenting their raw data, so that accuracies and other measures could be reconstructed. They discussed the variations in accuracies of phlebography and the usefulness and complications of metrizamide. They outlined a precise range of indications for phlebography.

Jepson et al. [24] studied two groups of patients, both undergoing surgery because of suspected lumbar disk lesions; one of these groups underwent myelography routinely and the other (1964-1968) did not. The authors argued that myelography did not much contribute to the indication for operation. In their discussion they reviewed this problem and pointed out that many articles with high accuracies are not applicable in clinical use. Their own data, which are complete, mentioned accuracies of 64-89% for myelography, depending on the strictness of the criteria. They did find fewer herniations in the group with myelograms than with clinical signs only (>90% in the latter), though this might reflect a relaxed threshold for operations.

Nelson and Gold [39] reported 10 patients with negative myelograms and positive operative findings. Clinical histories and a detailed tabulation of data are included. Depending on the criteria, all 10 or eight of the 10 patients had a correct CT finding. However the populations was heavily biased against myelography, and there was no indication about the selection as to considering all similar cases.

Raskin and Keating [40] addressed the problem of whether CT or myelography should be performed first. In this study with both metrizamide and Pantopaque myelograms, 106 patients had both procedures within 6 weeks (apparently for sciatica). Thirty-nine percent had surgery; in the other patients CT was compared with myelography. There is a good description of CT criteria for herniated disk. They explained that, due to the different way of visualizing, CT and myelography will not provide exactly the same results (as to precise localization), but often will give the same conclusions as to operation. They made a distinction between major and minor discrepancies, but had a rather large number of inconclusive investigations. As for myelography, they argued, that most of the inconclusive cases were caused by anatomic variations. We believe these should be counted in the category "major discrepancies". They gave rather low accuracies, which was caused partially by applying strict criteria, but also by the low prevalence of herniation in their group. However, their data did not allow a complete construction.

Stoeter et al. [41] verified CT findings in 106 patients by either surgical or CT findings, using a rather large doubt category and plural diagnoses (stenosis and other osseous causes, postoperative status). Consequently, their data were not comparable. There was no distinction between correctly positive or negative findings. Their moderate accuracy (74%) likely could be explained by the rather strict criteria applied. Attention was given to localization (medial vs lateral). They offered a number of practical considerations as to when to use CT or myelography as the first investigation.

Tchang et al. [42] analyzed 52 patients, with surgically confirmed disk herniations; 45 of them also had myelography. They make a distinction between original interpretation (actually used for patient management), interpretation with the use of clinical knowledge and other tests, and blinded evaluation. In their classification, the patients apparently had only a single abnormality. Their data were complete; however their selected population did not contain operations with negative findings or patients without herniations. Their accuracies were 94% and 87% for CT and myelography, respectively. They described their technique and discussed the differential diagnosis of CT findings.

Also in 1983, Griebel et al. [43] reviewed 100 patients who underwent surgery because of herniated disks; all had CT scans. The authors found better results for CT than for myelography, but did not mention the role of either in the decision to operate. Their classification of three cases as neither correct, false positive, or false negative, but as misinterpreted by the radiologist, is somewhat unusual. Their data did not allow a reconstruction.

Moufarrij et al. [26] selected 50 patients who underwent surgery for herniated disks, with exclusion of previously operated patients and those with suspected stenosis. CT findings were available for all patients and myelographic findings were available for 46. The patients were divided into five categories based on CT

diagnosis, including spondylosis and stenosis. From their four-way table (CT vs myelography vs surgical findings vs outcome), the test qualities were reconstructed by the reviewers. CT appeared more accurate if the only finding was herniation.

The 1983 article of Pythinen et al. [44] reviewed a series of 214 patients who (apparently) had both myelography and CT (in this order) for different indications (177 for herniated disk). They found that CT gave additional information in 12%, and in 7% (16 patients) the therapy was changed by CT. Detailed data were provided about these 16 cases, but there was no follow-up of the non-operated cases (61%). Their data did not allow a reconstruction.

In 1984, Bell et al. [27] reviewed the CT and myelographic findings in 122 patients with surgically proved pathology of herniated lumbar disk or stenosis or both. In this excellent article the herniated disk and the stenosis were considered for the first time together in a numerical evaluation of radiographic procedures, and the idea was developed of a local and global diagnosis. The images were evaluated blindly, which probably gave a lower accuracy. There was a range of accuracies, depending on strict or relaxed criteria. The literature review was extensive. It was not mentioned whether there were patients with the same diagnosis who had to be excluded because neither CT nor myelographic findings were available. The indications for operations were not mentioned.

In another 1984 study, Bosacco et al. [28] described 134 patients with suspected lumbar disk herniations (excluding patients with previous spinal surgery or likely stenosis). The CT and myelographic findings in these patients were evaluated with clinical information, but separately. For 61% conservative treatment was chosen, reflecting either a different patient selection or a different indication for surgery from most of the other studies. For these patients only a correlation between CT and myelography was presented, without specifying the fraction of false-positives and negatives. For 52 patients, the results of CT and myelography were compared with surgical findings. The data were complete for reconstruction of a two-by-two table. The study provided a description of the clinical symptoms and related them to CT, myelographic and surgical findings. The important role of workmens' compensation in negative cases was illustrated. The relative merits of the procedure was discussed.

Also in 1984, Valat et al. [29] presented phlebographic data from 104 operated patients in comparison with surgical findings, in both a per patient and per level analysis (complete data). If clinical presentations and plain radiographs were consistent as to level and site, phlebography did not add much information.

In 1985, Kampmann et al. [30] analyzed the data of 158 patients with sciatica in light of a number of questions: CT and myelography were compared for 36 nonoperated patients (90% correspondence); for the operated patients the CT (and part of the myelographic) diagnoses were compared with surgical findings. The criteria for correct diagnoses were compared with surgical findings. The

criteria for correct diagnosis were rather strict: the analysis was level-specific, but no errors were found in level or side of localization. The erroneous radiologic diagnoses were discussed, and the authors maintained that a myelogram would not have helped overcome the shortcomings of CT (based on three cases); they concluded that myelography is indicated if clinical signs and CT disagree.

In 1986, Greenough et al. [45] followed 22 patients who had had CT because of clinically suspected lumbar disk herniations with negative or equivocal myelograms. Although the prospective approach was completed for all patients, the numbers were small and the selection bias (against myelography) clear, so we did not include this study in Table 1. It is interesting to note that only eight of the 12 CT positive patients had surgery.

Also in 1986, Kratzer and Hipp [46] analyzed the CT and myelographic findings of 133 patients with a clinical suspicion of a herniated disk, of whom 93 had surgery. The relative advantages of CT and myelography were discussed in detail, with special attention to discrepancies between the clinical presentation and the radiologic findings, including recurrent herniation. The numeric data did not allow reconstruction of a two-by-two table.

In another 1986 article, Modic et al. [31] compared the MR, CT and myelographic findings (without clinical information) in 60 patients with suspected lumbar disk herniation or spinal stenosis, but without prior known diagnosis or imaging results. Forty-eight of these patients had surgery, five were normal, and seven had other disease. There were no negative explorations, so the test qualities in this review are not applicable. The lack of negative explorations suggests a rather high threshold for operation or for referral to the radiologist, but this was not confirmed. The fact that all normal MR studies were also normal on CT and myelography contradicts this assumption. Because the criteria for correlation of the radiologic procedures were strict (two diagnoses, level-specific stenosis), there were still discrepancies between the procedures, which were treated and commented on for the various combinations. In brief, the agreement of CT and myelography with MR was 87% and agreement with surgical findings was 83, 83, and 72% for MR, CT, and myelography, respectively. The relative merits of the procedures were discussed in detail. The authors concluded that MR (with surface coil) is equal to myelography or CT and that no single procedure alone is adequate.

In 1987, Schoedinger [47] described the CT, myelographic, phlebographic, and electromyelographic findings in 100 patients with surgically verified lumbar disk herniations. The various combinations of these tests have accuracies between 56 and 100%, but no single procedure predominated. For the most part, discrepancies in levels were disregarded. The possibility of false-positive results was mentioned only in the last section of the article, and there were not enough data to reconstruct the single test qualities.

Also in 1987, Voelker et al. [48] analyzed the metrizamide CT scans and myelograms of 80 patients who had either lumbar disk herniation or spondylosis at surgery. The images were evaluated blindly, which, together with the two surgical diagnoses under consideration, largely eliminated circular selection bias. The difference in accuracy between a per patient and per level analysis was explored. Lumbar disk herniation and spondylolysis did not differ in detectability between CT and myelography. Patients who had undergone surgery previously were shown to be measurably more difficult to diagnose. Their data did not allow a reconstruction of the test qualities, because the authors did not mention the existence or non-existence of false-positive diagnoses and seemed to be aiming only at not missing a diagnosis. This also affected their treatment of combined use of the tests.

Analysis

The aspects described above are presented in Table 1 for those articles that contained data allowing calculation of sensitivity and specificity, or that contained otherwise complete numeric analyses. In the discussion that follows, the findings in the 25 articles listed in Table 1 are summarized.

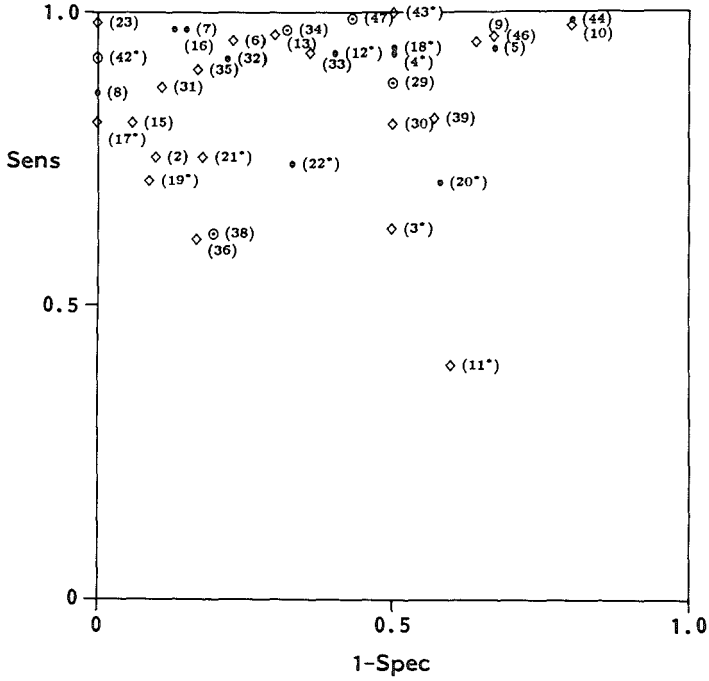
The selection criterion was operated patients in 10 articles, operated patients with supplemental data in four articles, surgically verified diagnosis of herniation or stenosis in four articles, clinical presentation in five articles, and unclear in two articles. The selection criteria *operated patients* and *surgically verified patients* represent a retrospective, rather than a prospective selection. The number of available observations (either levels or patients) was less than 50 in seven articles, between 50 and 150 in 13, and more than 150 in five. Only one article had a prospective selection criterion and more than 150 observations. Four papers had blind interpretation, two with a clinical interpretation, and two with both types of interpretation on the same population. (The complement to Haughton et al. [23] is presented in a separate article [37] that is not included in Table 1.) The other 15 papers did not state whether they used clinical information. Four papers did apply more than one threshold for the diagnostic criterion.

Myelography and surgery were compared in 19 articles and phlebography and surgery in nine. CT and surgery were compared in seven articles.

ROC analysis of data

For those articles that contained a comparison of phlebography, CT, or myelography with surgery, the sensitivity and specificity, which are shown in Table 1, are also presented in a Receiver Operating Characteristic (ROC) space (Figs. 1 and 2) [49].

Fig. 1 — Test qualities from sensitivity (Sens) and specificity (Spec) of phlebography (*solid circles*), myelography (*diamonds*), and CT (*circles with dots*). Sensitivity (1 – true-positive ratio) is on Y axis; specificity is on X axis. Numbers in parentheses correspond to 25 consecutive references listed in Table 1. Those with an asterisk denote studies with a marked bias or other reasons for being out of order.

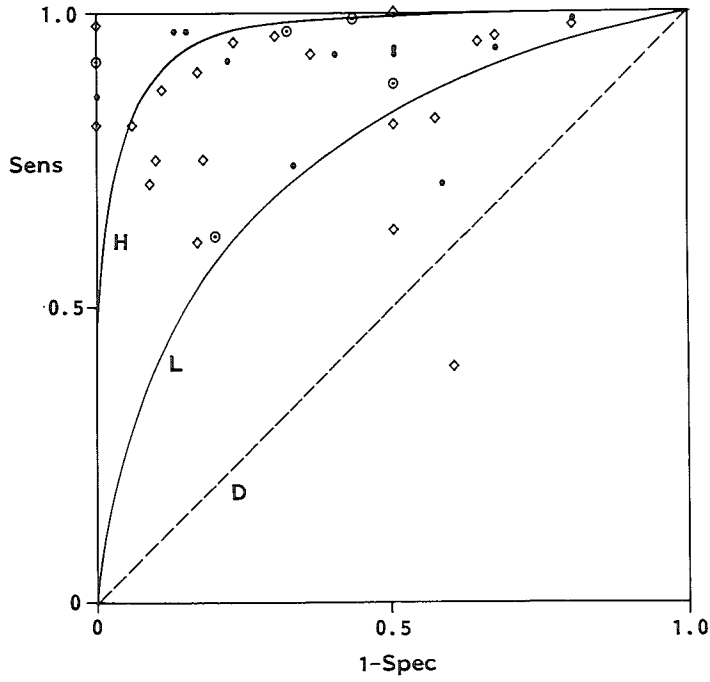


This allows us to see the difference (between studies) in both sensitivity and specificity, which must be always considered together.

The ROC space is divided into three regions (Fig.2) to distinguish between relatively well and poor performing tests. The positions of the boundaries are rather arbitrary. The best and worst performing sixth parts are pragmatically called deviant. Though the number of observations in general is too small for individual points (studies) to be compared, we can look at the total pattern that emerges by studying the extremes.

Lower quality outliers. — Roland et al. [15] reported myelographic findings in "patients with a clinically suspected disk but negative or equivocal myelograms". Theirs is the only study with results that are below the diagonal D in Figure 2, which means that the test outcomes should be interchanged for optimal results. Gargano et al. [6] used the same biased selection criterion as Roland et al. In their fourth group, Lotz et al. [17] reported phlebographic results from a "blinded" experiment; the results were manipulated by the reviewers, in that the quite large "dubious" category was distributed over the cells according to the marginals. This

Fig. 2 — Test qualites from sensitivity (Sens) and specificity (Spec) of phlebography (*solid circles*), myelography (*diamonds*), and CT (*circles with dots*). Sensitivity (1 – true-positive ratio) is on Y axis; specificity is on X axis. Curves and broken line connect sensitivity/specificity pairs of tests with a constant quality, but with different cutoff points between normal and abnormal. Diagonal line (D) represents a completely uninformative test; Low (L) and High- (H) quality tests are represented by H and L curves.



degrades the results. The second comparison of Jepson et al. [24] is a variant with stricter criteria; their first study is in the medium region. Moufarrij et al. [26] and Claussen et al. [21] do not provide clear explanations for their (relatively) poor results.

High quality outliers. — Thijssen et al. [18] considered only myelograms of "adequate quality", which renders the test results overly optimistic. The second comparison of MacNab et al. [11] was limited, because patients without prior surgery were studied; in addition the evaluation was blinded, making interpretation more difficult. In the studies of Miller [12] and Meyenhorst [8] (second and third comparisons), the selection criteria of the patients were not described in enough detail to attribute or deny any influence from it. The results in the third comparison of Bosacco et al. [28] were quite possibly affected by artifacts, since specificity was only based on only two cases.

If we exclude all articles with an apparent bias, and the article of Lotz et al. [17] (because of the manipulation by the reviewers), the remaining results show, despite their large differences in sensitivity or specificity, considerable coherence when regarded as a ROC curve. (Excluded articles are marked with an asterisk in Figure 1.)

In Table 1, it is hard to conclude that myelography is better or worse than phlebography. Only MacNab et al. [11], Meyenhorst [8], and Lotz et al. [17] provided direct comparisons between the two procedures on the same population and without apparent bias toward one. MacNab et al. and Meyenhorst found phlebography to be superior to myelography; in the Lotz et al. article, myelography was equal to or better than phlebography.

Discussion

Ever since its introduction in 1936 [50], myelography had been widely used for the detection of lumbar disk herniation. The introduction of alternative diagnostic tools raised the question of whether myelography could be replaced by one of the newer methods and in what cases which procedures should be chosen.

Various studies in which different radiological techniques are compared have tried to answer these questions. The conclusions of these studies, however, are sometimes conflicting. Because of this, the need became obvious for a review and analysis of the literature that tried to explain these discrepancies.

In the current review, we screened the available literature addressing the capabilities of myelography, epidural phlebography, and CT to detect lumbar disk herniation and spinal stenosis. Phlebography seems to have lost its place in daily practice with the introduction of CT (it was used especially for lateral and intraforaminal herniations). It is included in our study because of its importance in the spectrum of diagnostic reliability and its role in the comparative studies reported in the literature. MR imaging is not included, because there too few comparative studies with enough patients.

There are of course subjective elements in the analysis of the reviewed literature. Though we tried to maintain uniformity in our analysis, many articles include special circumstances or boundary-case presentations. Sometimes a different approach to what is clinically relevant prevented us from using data that were otherwise valid. It was not our intention to judge the quality of single articles, and, as has been mentioned, other qualities of these articles were overlooked.

Despite the necessarily subjective classification, most of the conclusions would remain the same if some papers were otherwise rated, because the pattern of this set of papers would not change. Though we tried to find all the relevant literature, it is possible that some data about myelography, CT, or phlebography are included

in other articles, lengthy reports, or monographs with another scope; if so, these were not referenced in the reviewed papers.

Our analysis and review concentrated on the capability of the radiographic procedures to predict surgical findings, not whether one or more of these procedures are required before deciding to operate.

Only some of the literature was suitable for interstudy comparisons. An important problem was the definition of the study population. Some studies had only a small number of patients; very often it was not even mentioned whether the images were evaluated with our without clinical information or to what degree the results of the images influenced the decision to operate. It remains unclear how the results were influenced by important factors such as mild or strict criteria for correspondence of radiologic and surgical findings. Too often, emphasis was on accuracy as a measure of the results, which gives less information than the simultaneous consideration of sensitivity and specificity in a ROC space.

The interpretation of the points in a ROC space should be treated with caution: In the regular ROC analysis it is assumed that there is a means of verifying the disease, independent of the test. Consequently, prevalence here refers to the detected prevalence in a clinical subpopulation. This cannot be maintained for the investigation under study. In most of the articles, the radiologic diagnosis will have played some role in the decision to operate. However (especially in the investigations with two radiologic procedures), they will not have *determined* the decision. This difference can be seen between groups 4 and 5 and groups 2 and 3 of Lotz et al. [17].

Whatever other arguments played a role in the decision to operate, and whatever weights may have been attached to the test results, together they appear to have placed the cutoff points of the majority of these papers along one ROC curve, if articles with apparent bias are excluded. One possible interpretation of this phenomenon is that the values assigned by patients and physicians to false-positive and false-negative outcomes vary among the cited studies and that the tests are of the same order of quality, despite the differences in study design and technique; this cannot be ascertained.

The influence of a positive radiologic procedure on the decision whether to operate will most likely cause the stated accuracy to be too optimistic, and this effect will be more pronounced for the medium or poor tests. Poor tests, however, will tend to have less influence on the decision to operate.

In Figures 1 and 2 we see that there is more variation in the specificity (0.2-1) than in the sensitivity (0.6-1). Apparently in most of the studies there was a tendency not to miss a possible herniation. Although a threshold was never mentioned in the original publications, the existence of implied values attributed to false-positive of false-negative diagnoses is undeniable from the set of articles.

A literature analysis always excluded the most recent publications, which are always behind the most recent developments. This is more meaningful for CT, where developments have been very fast, than for myelography or phlebography. It is possible that if this analysis were repeated after a few years, CT would have a stronger position. However, based upon the findings of the literature reviewed so far, there are no clear indications to consider myelography or CT or phlebography superior over the other studies in the diagnosis of lumbar herniated disk.

References

1 Di Chiro, G. Computed tomography of the spinal cord after lumbar intrathecal introduction of metrizamide (computer-assisted myelography). *Radiology* 1976; **120**: 101-104.

2 Coin CG, Chan YS, Keranen V, et al. Computer assisted myelography in disk disease. *J Comput Assist Tomogr* 1977; **1**: 398-404

3 Helander CG, Lindblom A. Sacrolumbar venography. *Acta radiologica* 1955; **44**: 410-416.

4 Perey O, Lind J, Wegelius C. Phlebography of the intervertebral plexus. *Acta orthop Scand* 1956; **25**: 228-233.

5 Nathan MH, Blum L. Evaluation of vertebral venography. *AJR* 1960; **83**: 1027-1033.

6 Gargano FP, Meyer JD, Sheldon JJ. Transfemoral ascending lumbar catheterisation of the epidural veins in lumbar disk disease. *Radiology* 1974; **111**: 329-336.

7 LePage JR. Transfemoral ascending lumbar cathetherization of the epidural veins. *Radiology* 1974; **111**: 337-339.

8 Meyenhorst GCH. Transfemoral epidural double-catheter venography in the diagnosis of lumbar disc herniation. [Thesis] Nijmegen, The Netherlands: Nijmegen Catholic University, 1979.

9 Hudgins WR. The predictive value of myelography in the diagnosis of ruptured lumbar discs. *J Neurosurgery* 1970; **32**: 152-162

10 Drasin GF, Daffner RH, Sexton RF, et al. Epidural venography: diagnosis of herniated lumbar intervertebral disc and other disease of the epidural space. *AJR* 1976; **126**: 1010-1016.

11 MacNab I, St. Louis EL, Grabias SL, et al. Selective ascending lumbosacral venography in the assessment of lumbar-disc herniation. *J Bone Joint Surg [Am]* 1976; **58**: 1093-1098.

12 Miller MH, Handel SF, Coan JD. Transfemoral lumbar epidural venography. *AJR* 1978; **126**: 1003-1009.

13 Mohsenipour I, Twerdy K, Pirker E. Eine Gegenüberstellung Röntgenbild — Operationsbefund bei lumbalen Bandscheibenerkrankungen. ROFO 1977; **127**: 540-543.

14 Moringlane JR, Voigt K, Seeger W. Vergleich myelographischer und intraoperativer Befunde beim lumbalen Bandscheibenvorfall. *Neurochirurgia (Stuttg)* 1977; **20**: 199-208.
15 Roland JR, Treil J, Larde D, et al. Lumbar phlebography in the diagnosis of disc herniations. *J Neurosurg* 1978; **49**: 544-550.
16 Cook PL, Wise K. A correlation of the surgical and radiculographic findings in lumbar disc herniation. *Clin Radiol* 1979; **30**: 671-682.
17 Lotz PR, Seeger JF, Gabrielsen TO. Prospective comparison of epidural venography and iophendylate myelography in the diagnosis of herniated lumbar disks. *Radiology* 1980; **134**: 127-132.
18 Thijssen HOM, Rombouts JJM, Walder HAD. Diagnostic accuracy of lumbar myelography in the detection of lumbar disk herniations. *Diagn Imaging* 1980; **49**: 188-192.
19 Gulati AN, Weinstein Z, Studdard W. CT scan of the spine for herniated discs. *Neuroradiology* 1981; **22**: 57-60.
20 Anand AK, Lee BCP. Plain and metrizamide CT of lumbar disk disease: comparison with myelography. *AJNR* 1982; **3**: 567-571.
21 Claussen C, Grumme T, Treisch J, Lochner B, Katzner E. Die Diagnostik des lumbalen Bandscheibenvorfalls. *ROFO* 1982; **136**: 1-8.
22 Fries JW, Abodeely DA, Vijungco JG, et al. Computed tomography of herniated and extruded nucleus pulposus. *J Comput Assist Tomogr* 1982; **6**: 874-887.
23 Haughton VM, Eldevik OP, Magnaes B, et al. A prospective comparison of computed tomography and myelography in the diagnosis of herniated lumbar disks. *Radiology* 1982; **142**: 103-110.
24 Jepson K, Nada A, Rymaszewski L.The role of radiculography in the management of lesions of the lumbar disc. *J Bone Joint Surgery [Br]* 1982; **64**: 405-408.
25 Sachsenheimer W, Hamer J, Müller HA. The value of spinal computed tomography in diagnosis of herniated lumbar discs. *Acta Neurochirurgica* 1982; **60**: 107-114
26 Moufarrij NA, Hardy Jr RW, Weinstein MA. Computed tomographic, myelographic and operative findings in patients with suspected herniated lumbar discs. *Neurosurgery* 1983; **12**: 184-188.
27 Bell GR, Rothman RH, Booth RE, et al. A study of computer assisted tomography. II. Comparison of metrizamide myelography and computed tomography in the diagnosis of herniated disk and spinal stenosis. *Spine* 1984; **9**: 552-556.
28 Bosacco SJ, Berman AT, Garbarino JL, Teplick JG, Peyster R. A comparison of CT scanning and myelography in the diagnosis of lumbar disc herniation. *Clin Orthop* 1984; **190**: 124-128.
29 Valat JP, Gatti P, Saindelle A. Phlébographie lombaire. Apport au diagnostic des lombosciatiques d'origine discale. *Sem Hôp Paris* 1984; **60**: 539-542.
30 Kampmann H, Schroedl P, Spranger M. Diagnostik lumbaler Bandscheibenvorfälle durch Computertomographie. Eine klinische Vergleichsstudie zwischen myelographischen, computertomographischen und operativen Untersuchungsergebnissen von 158 Patienten. *Röntgenblätter* 1985; **38**: 387-391.

31 Modic MT, Masaryk T, Boumphrey F, Goormastic M, Bell G. Lumbar herniated disk disease and canal stenosis: Prospective evaluation by surface coil MR, CT, and myelography. *AJNR* 1986; **7**: 709-717, *AJR* 1986; **147**: 757-765.
32 Gershater R, St Louis E. Lumbar epidural venography. Radiology 1979; 131: 409-421.
33 Williams AL, Haughton VM, Syvertsen A. Computed tomography in the diagnosis of herniated nucleus pulposus. *Radiology* 1980; **135**: 95-99.
34 Hanson RD. Computed tomography in the diagnosis of lumbosacral disc herniation. *Ariz Med* 1981; **38**: 839-842.
35 Harley WD, Sava GF, Fleming RJ. Computerized tomography in the evaluation of lumbar disc herniation. *Conn Med* 1981; **45**: 349-352.
36 Dublin AB, McGahan JP, Reid MH. The value of computed tomograhic metrizamide myeolography in the neuroradiological evaluation of the spine. *Radiology* 1983; **146**: 79-86.
37 Eldevik OP, Dugstad G, Orrison WW, et al. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. *Radiology* 1982; **145**: 85-89.
38 Herkowitz HN, Wiesel SW, Booth RE, et al. Metrizamide myelography and epidural venography. Their role in the diagnosis of lumbar disc herniation and spinal stenosis. *Spine* 1982; **7**: 55-64.
39 Nelson MJ, Gold LH. CT evaluation of intervertebral foramina lesions with normal or non diagnostic myelograms. Report of ten cases. *Comput Radiol* 1983; **7**: 155-160.
40 Raskin SP, Keating JW. Recognition of lumbar disk disease: Comparison of myelography and computed tomography. *AJR* 1982; **139**: 349-355.
41 Stoeter P, Schneider I, Bergleiter R, Ebeling U. Diagnostischer Wert der computertomographischer Untersuchung der Lumbosakralregion bei Patienten mit Lumboischialgien. *ROFO* 1982; **136**: 515-524.
42 Tchang SPK, Howie JL, Kirkaldy-Willis WHK, et al. Computed tomography versus myelography in diagnosis of lumbar disc herniation. *J Can Ass Radiol* 1982; **33**: 15-20.
43 Griebel R, Tchang S, Khan M, et al. Correlation of computed tomography with surgical diagnosis in lumbar disc disease. *Can J Neurol Sci* 1983; **10**: 248-251.
44 Pythinen J, Lahde S, Tamska EL, et al. Computed tomography after lumbar myelography in lower back and extremity pain syndromes. *Diagn Imaging* 1983; **52**: 19-22.
45 Greenough CG, Dimmock S, Edwards D, Ransford AO, Bentley G. The role of computerised tomography in intervertebral disc prolapse. *J Bone Joint Surg [Br]* 1986; **68**: 729-733.
46 Kratzer M, Hipp E. Stellenwert der Myelographie und Computertomographie bei Problemfällen in der lumbalen Bandscheibendiagnostik. *Z Orthop* 1986; **124**: 107-111.
47 Schoedinger GR III. Correlation of standard diagnostic studies with surgically proven lumbar disk rupture. *South Med J* 1987; **80**: 44-46.

48 Voelker JL, Mealey J Jr, Eskridge JM, Gilmor RL. Metrizamide-enhanced computed tomography as an adjunct to metrizamide myelography in the evaluation of lumbar disc herniation and spondylosis. *Neurosurgery* 1987; **20**: 379-384.
49 Metz CE. Basic principles of ROC analysis. *Sem in Nucl Med* 1978; **8**: 283-298.
50 Hampton AO, Robinson JM. Roentgenographic demonstration of rupture of the intravertrbral disk into the spinal canal after injection of Lipiodol. *AJR* 1936; **36**: 782-803.

3

COMPARATIVE DIAGNOSTIC PERFORMANCE OF THREE RADIOLOGICAL PROCEDURES FOR THE DETECTION OF LUMBAR DISK HERNIATION

Jan WPF Kardaun, MD [1]
Otto JWF Kardaun, PhD [2]

[1] Dept Health Statistics, Central Bureau of Statistics
Voorburg, The Netherlands

[2] Mathematical Institute, Groningen University,
Groningen, The Netherlands.
presently: Max Planck Institute for Plasma Physics,
Garching, Federal Republic of Germany.

The views and assertions herein are those of the authors. In particular, they are not to be construed as official or reflecting the views of the Netherlands Central Bureau of Statistics.

Abstract

Literature data on the diagnostic performance of phlebography, myelography, and CT scan applied to patients with suspected lumbar disk herniation (LDH), are analyzed to extract maximal information about their relative discriminatory power. Seventeen papers meeting the selection criteria contain 13 reports on myelography, 6 on phlebography, and 5 on CT. Sensitivity and specificity are considered simultaneously in logistic ROC space. The reports of each procedure are effectively summarized by a linear regression in logistic ROC space. Taking into account the individual confidence regions of sensitivity and specificity obtained from each report, the slope of the regression line is estimated by Generalized Least Squares (ML). This approach also allows to test the assumption of a common odds ratio (i.e., of a unit slope). The simply to determine common odds ratio as well as the perpendicular distance between the origin and the regression line (as a good approximation to the area under the ROC curve) are used as a measure for the discriminatory power of the procedures. For CT, homogeneity of sensitivity turns out to be much more likely than a common odds ratio. Based on the available, retrospective data, *phlebography* appears to have the highest performance in visualising an LDH, followed by *myelography* and *CT*.

Keywords: lumbar disk herniation / myelography / phlebography / CT / ROC analysis / diagnostic test evaluation

In press, Methods of Information in Medicine

Introduction

Lumbar Disk Herniation (LDH) is a rather common disease: the number of hospital admissions for LDH in The Netherlands is estimated to be approximately two per thousand persons per year [1]. This disease causes considerable pain, discomfort and loss of working ability to the patients [2]. Myelography has been widely used for the visualisation of lumbar disk herniation since its introduction in 1936 [3]. It is an X-ray procedure of the (lumbar part of) the spinal canal, with the injection of X-ray contrast into the dural enveloppe that surrounds the myelum. Several other radiological procedures, with or without contrast injection, have also been applied. Phlebography consists of an X-ray procedure with contrast injection by catheterisation into the plexus venosus vertebralis. Contrast is usually not applied for a CT scan of the lumbar spinal canal. The latest technology is Magnetic Resonance Imaging (MRI), but at the moment it is (in Western Europe) available in major hospitals only. The question as to which of these procedures is superior to the others in diagnosing an LDH remains under discussion.

In a recent literature review [4], relevant studies since 1970 which presented data on the ability of phlebography, myelography or CT scan to visualize an LDH were summarized and commented. This paper intends to give a statistical analysis of the sometimes contradictory results from these reviewed studies. (Such integration of results from a number of individual studies is sometimes termed *meta-analysis*, see [5, 6].) It will be clear that by combining the results from many studies, the effective sample size is increased. On the other hand, differences in study design or a tendency towards a more specific or more sensitive result of the individual studies have to be accounted for.

Data and Methods

Reports in regular journals since 1970 on the diagnostic capabilities of phlebography, myelography or CT scan for lumbar disk herniation were screened on the following criteria:

(1) A minimum study population of 20 patients,
(2) the availability of sufficient data in tables or text to describe the results in a 2 × 2 table, consisting of LDH visualization or no visualization versus confirmation or denial of an LDH by surgery (reasons for failing this criterion were an incomplete presentation of the data in 11 reports, and the selection of only patients with a surgically verified LDH in 6 reports),
(3) no clear bias against one of the procedures (some studies for instance selected only patients for phlebography after a negative myelography).

Of the 39 relevant papers reviewed in [4], 17 met these criteria. One paper was excluded, because the authors used a rather large doubt category which made the results incomparable with other papers [7]. One recent paper [8] was added. For the resulting 17 papers the relevant data, displayed in Table 1, are year of publication, name of the first author, the procedure(s) involved and four numbers

Table 1 — Summary of the test qualities of phlebography (P), myelography(M) and CT scan (C) for the detection of lumbar disk herniation from selected literature since 1970.

first author	ref nr	year publ	radiol. procedure	sample size	prev	sens	spec
Hudgins	10	1970	M	135	0.79	0.75	0.90
Drasin	11	1976	P	19	0.84	0.94	0.33
Macnab	12	1976	M	50	0.74	0.95	0.77
			P	50	0.74	0.97	0.85
Miller	13	1978	P	38	0.92	0.86	1.00
Mohsenipour	14	1977	M	401	0.93	0.95	0.36
Moringlane	15	1977	M	140	0.96	0.98	0.20
Cook	16	1979	M	62	0.84	0.96	0.70
Meyenhorst	17	1979	M	151	0.52	0.81	0.94
			P	151	0.52	0.97	0.87
Thijssen*	18	1980	M	104	0.91	0.98	1.00
Claussen	19	1982	M	23	0.91	0.81	0.50
			C	26	0.92	0.88	0.50
Fries	20	1982	M	192	0.91	0.87	0.89
			P	227	0.90	0.92	0.78
Haughton	21	1982	M	55	0.55	0.93	0.64
			C	55	0.55	0.97	0.68
Jepson	22	1982	M	55	0.89	0.90	0.83
Moufarrij	23	1983	M	46	0.85	0.82	0.43
			C	50	0.90	0.62	0.80
Valat	24	1984	P	104	0.95	0.99	0.20
Kampmann	25	1985	M	31	0.90	0.96	0.33
			C	122	0.91	0.99	0.57
Schipper	9	1987	M	263	0.87	0.83	0.71
			C	235	0.84	0.71	0.79

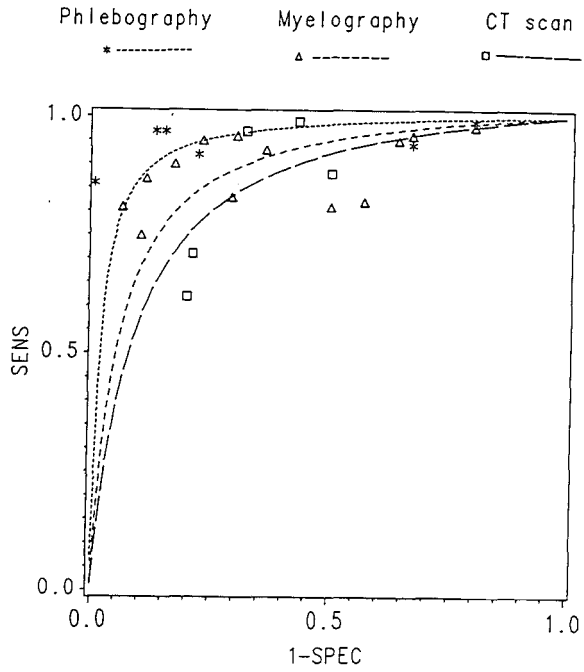
* This observation is considered as an outlier and is not included in the analyses.
prev = prevalence; sens = sensitivity; spec = specificity.

characterizing the 2 × 2 table: sample size, LDH prevalence (PREV), sensitivity (SENS), and specificity (SPEC).

Representation in ROC Space

The most widely used measure of the test quality in the reviewed clinical papers is the "accuracy" (or 1 – error rate), which is the ratio of the number of correct X-ray diagnoses and the total number of X-ray diagnoses. Obviously, this measure provides only limited insight in differences in quality between the tests. A better approach is to consider jointly the sensitivity and specificity of the tests in a so-called ROC space [25]-[27]. In Figure 1 the test qualities of phlebography, CT and myelography are shown in the traditional ROC-space representation, with sensitivity (the probability that the patient has a positive test result, given that he has the disease) ranging from 0 to 1 on the y-axis and specificity (the probability that the patient has a negative test result, given that he does not have the disease)

Figure 1 – Phlebography, myelography, and CT scan in standard ROC space



ranging from 1 to 0 on the x-axis. The dotted lines indicate curves of "equal test quality", where the observer may vary his classification threshold, depending on whether he prefers a high specificity or a high sensitivity.

Most of the reports seem to have results scattering reasonably close around a fitted line. To assess this visual impression, it would be informative to see the confidence regions for the sensitivity and specificity of individual reports. In the traditional ROC space it is rather complicated to calculate these exactly. Tables and nomograms [28] give only one-dimensional confidence intervals, no confidence regions. Normal approximations to the binomial distribution, fall short when the (sensitivity, specificity) coordinates come close to the axes. Some methods are discussed in Koopman [29].

Logistic ROC Space

If the logit transformations of sensitivity and specificity are used to form the ROC space, the asymptotic variances are easily calculated by a simple formula. The normal approximation on the estimated sensitivity and coordinates on the logistic scale is expected to be somewhat better than on the traditional (0 — 1) scale for the region close to the axes.

In this normal approximation, confidence regions are easily constructed (for the notation of a 2 × 2 table used in this paper, see the Appendix):

$$\text{Var}(\text{logit}(\text{sens})) = (1/a + 1/b) \text{ and}$$

$$\text{Var}(\text{logit}(1\text{–spec})) = (1/c + 1/d)$$

The logistically transformed scales can be interpreted as follows. If the feature distributions of the X-ray images of normals and diseased, condensed into one-dimensional characteristics, are logistic distributions with equal variances, then the line of constant test quality becomes a straight line with a 45° angle to the axes. The abscissa of this line, equal in magnitude to the ordinate, is in that case a measure of the discriminatory power of the X-ray procedure. It is equal to the distance of the means of the two logistic distributions of the X-ray characteristics of normals and diseased (for comparison, see [30], chapter 4). If we consider two arbitrary logistic distributions, the line of constant test quality has an arbitrary slope β, where β^2 equals the ratio of their variances.

In ROC analysis the distribution of the features is often assumed normal, in which case a probit transformation would be appropriate. However, logistic and normal distributions are often hard to discriminate in practice [31, 32]. Hence one can use both approximations to a "real" situation.

The logistic transformation allows for an additional interpretation which does not use the concept of a probability distribution of the features. After a logistic transformation, the line with unit slope can also be interpreted as the line of

constant odds ratio ($(a/b)/(c/d)$) for a series of 2 × 2 tables. The odds ratio, a well known measure in epidemiology [33, 34], is a useful one-dimensional summary of the quality of a procedure and it can be easily estimated.

Our data indicate that sometimes a straight line should be fitted that does not run under a 45° angle with the axes. A derivation of how to perform this routinely, and how to calculate confidence intervals for its slope is given in the Appendix.

Results

In Figure 2 are shown for phlebography, myelography, and CT scan, respectively, the estimated (sensitivity, specificity) pairs of the individual reports in a logistic ROC space. Ellipses indicate the confidence regions of the sensitivity and specificity. The half-axes are equal to 1 standard deviation, which amounts to approximately a 40% confidence area. In Table 2 the estimated common odds ratios, together with approxmimately 95% confidence intervals and Breslow-Day statistics [35] for the homogeneity of the odds ratios are shown.

Phlebography

For the phlebography reports (Fig. 2a), it is clear that most of them have a higher sensitivity than specificity. If there is a difference in the variance between sensitivity and specificity then that of specificity appears to be larger. Three or four of the six reports have a larger confidence area than the reports of Meyenhorst [16] and

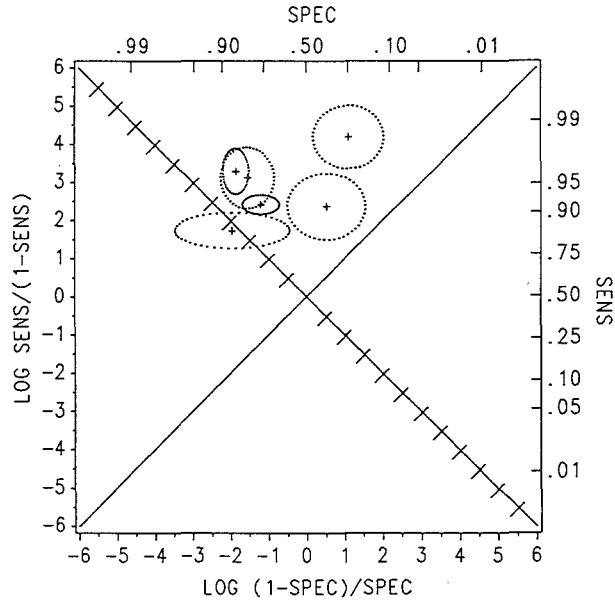
Table 2 — Calculation of a common odds ratio (Mantel & Haenszel) and testing for homogeneity of the odds ratios (Breslow & Day), and calculation and testing of a linear fit, for each of the radiological procedures.

Procedure	Total sample size	Common O.R.	C.I.	Breslow-Day P	β	C.I.	α	C.I.	*SS*	d.f.	P
Phlebography	601	62	36-106	0.17	0.50	0.02 - 1.6	3.3	2.5 - 5.0	5.2	4	0.40
Myelography	1630	19	14- 27	0.11	0.67	0.4 - 1.1	2.6	2.2 - 3.2	14.5	11	0.20
CT scan	498	12	7- 21	0.08	4.0	1.4 - –11	5.5	2.7 - –6.9	0.8	3	0.85

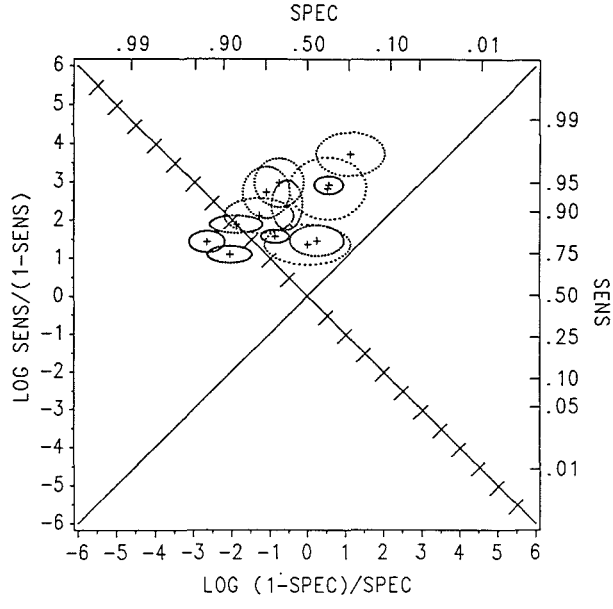
The regression equation is $\log(\text{sens}/(1-\text{sens})) = \alpha + \beta(\log((1-\text{spec})/\text{spec}))$. C.I. denotes a 95% confidence interval for the pertaining regression parameters, *SS* denotes the residual sum of squares from the linear fit, and P the corresponding approximate P-value under the hypothesis that a linear fit holds.

Figure 2 – Probability contour (40%) of sensitivity and specificity in logistic ROC space

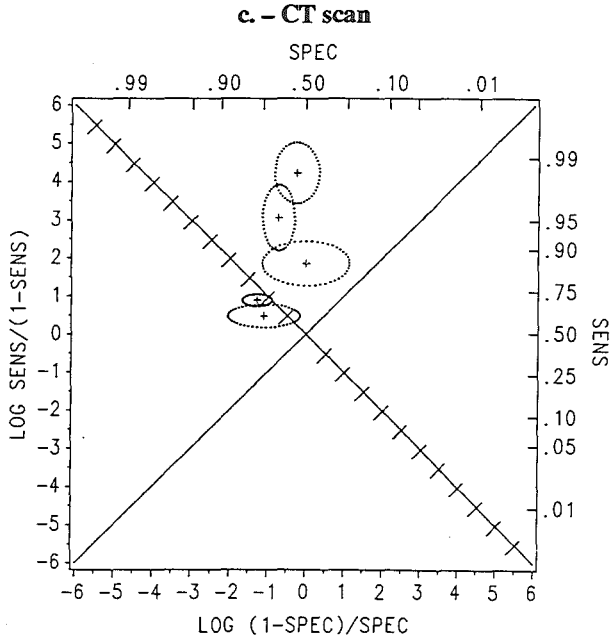
a. – Phlebography



b. – Myelography



(Fig. 2 continued)



Fries et al. [19], which have larger patient samples. It is remarked that the area of the ellipses is inversely proportional to the sample size and it also depends on sensitivity, specificity and prevalence. Though the best fitted straight line has a slope of 0.5, a unit slope cannot be rejected, both by the C.I. for β and by the Breslow-Day statistic (Table 2). It is noted, however, that, partly due to the small number of reports, the slope is ill-determined. The hypothesis of a straight line is not rejected. (P=0.40 by an approximate χ^2_{4df} test.) The estimated common odds ratio (MH:62) and the estimated odds ratio at specificity 0.5, i.e. $\exp(\hat{\alpha})$ (KK:27) are quite large (Table 2).

Myelography

One of the reports on myelography [17] seems clearly an outlier, and was excluded from our analyses (Fig. 2b). The results in their paper are extremely good, but this may be well due to a biased selection of myelograms. This bias is suggested by the fact that the authors argued that myelography could be equal to phlebography, provided only myelograms of good quality were considered.

Most of the other reports show relatively small confidence areas, and a linear fit seems adequate (P=0.20). The estimated slope (β) does just reach unity in its upper confidence limit, while the Breslow-Day statistic, which is less sensitive against the specific alternative of a straight line with unit slope, is more comfortable about a common odds ratio (P=0.11). The estimate of the common odds ratio (MH:19) and the common odds ratio at specificity 0.5 (KK:13.5) is less than that of phlebography, but still considerable.

CT Scan

Only one report [8] has a reasonable small confidence area (Fig. 2c). The other reports have rather small sample sizes. Though a linear fit seems appropriate (P=0.85), a unit slope is unlikely as is suggested by Figure 2c and can be seen from the C.I. of β. Notice that a straightforward application of the Breslow-Day statistic would falsely suggest that a common odds ratio for the CT scan reports cannot be rejected. An infinite slope of the regression line is included in the C.I. of β. Note that the slope interval of β, from 1.4 to –11, corresponds to an angle of 54 to 95 degrees. From Table 2 one can see that the odds ratio at specificity 0.5 ($\exp(\hat{\alpha})$) is 245 but, because of the large slope β, the odds ratio is strongly dependent on the specificity (or sensitivity) chosen. At sensitivity = 0.75, for instance, the odds ratio is 0.97.

Phlebography, Myelography and CT compared

There are several constraints to this dataset and our approach — which will be discussed later — but based upon the present analysis, we can compare the procedures in the following ways. Using the MH odds ratio, phlebography appears better than myelography; their C.I.'s do not overlap. A common odds ratio for CT does not seem adequate, given the slope of the fitted regression line. Considering odds ratios at specificity 0.5 ($\exp(\hat{\alpha})$), it appears that phlebography is somewhat better than myelography, and worse than CT, but all their C.I.'s do overlap. Moreover, the restriction of comparing at specificity 0.5 may not always be adequate.

An additional way of comparing test qualities in ROC space is comparing the area under the ROC curve ([25, 26]). Using estimates from Table 2, the area under the ROC curve was calculated as 94% for phlebography, 88% for myelography, and 79% for CT. Analogous to the situation in the normal ROC space [25] a good substitute for the area under the ROC curve is the distance δ from the origin to the fitted regression line. For phlebography δ = 2.9 (2.4—3.4), for myelography δ = 2.2 (1.8—2.4), and for CT δ = 1.4 (0.5—2.1). (The numbers between parentheses are 95% C.I.'s.) If differences between these procedures are tested at

a 5% level, assuming independent observations and an approximate normal distribution for δ, the difference between phlebography and myelography is significant (Z=2.40), between myelography and CT is non-significant (Z=1.87), and between phlebography and CT is significant (Z=3.18). These conclusions are not modified if some dependence between the observations of CT and myelography is assumed. The difference between phlebography and myelography becomes borderline significant (5%) if a correlation coefficient between the estimated δ values of 0.6 is assumed. (It is noted that the patient groups undergoing CT or phlebography were independent.) The main conclusions are again not altered if, by Bonferroni's inequality, the individual tests are carried out at 5%/3=1.7% significance level (Z=2.1), to accomodate for the multiple testing.

Discussion

Validity of the Data

The present analysis tries to bring some order in the multitude of the sometimes conflicting results from reports on the diagnostic qualities of three radiological procedures for lumbar disk herniation. The use of data from the published literature (meta-analysis) allows to use a larger number of patients than can be encompassed by a single study or than can be reasonably subjected to randomization. However, as the single reports collected their data independently, biases are introduced by differences in patient selection, radiological technique, diagnostic criteria and tabulation of the results [36]. These aspects of the test results are usually hard to quantify and not always given full attention in clinical reports. Three of the, in our opinion, most important sources of bias concerning the question posed will be discussed. It should be clear, however, that these sources of bias also apply to single study analyses, and that as yet no practical solution has been found. The nature of the research question imposes severe constraints on the possibilities for collecting data, as there are undesirable side-effects of radiological investigation and surgery. These constraints are:

a. The selection criteria of patients for the study — more concretely of those patients whose complaints point to an LDH and warrant a phlebography, myelography or CT scan — have an influence on the estimated prevalence of LDH's in the study population and also on the odds ratio and its C.I. Ideally, differences in such selection criteria between the various investigations should be avoided.

Moreover, a further selection of patients for a specific diagnostic procedure may bias the comparison between the procedures. From this point of view, it has to be avoided that the difficult cases are reserved for a new procedure, at times scarce, such as CT scan, or a more elaborate procedure, such as phlebography. Even

though in this meta-analysis we excluded reports with a clear preferential selection for one of the procedures, it cannot be concluded that all other reports will have had an unbiased selection of patients for each procedure, especially as comparative papers with an explicitly unbiased study design are scarce [7, 20].

b. The operative verification of a presumed LDH is not perfect. Even though operative verification is treated in the clinical literature as a gold standard, it is agreed upon that this standard is sometimes fallible. Hence, the term "gold-plated standard" could be coined. Two surgeons may disagree on a borderline herniation; an LDH, in a very lateral location for instance, can be missed during surgery. However, no better way of verification is as yet possible. An additional limitation of this standard is that a "gold-plated" diagnosis is not available for those patients that are not operated upon. A verification by functional improvement, which would be available for all patients, is notably hard [37].

c. The decision to operate is primarily based upon the presumed absence or presence of an LDH. Therefore, the result of the radiological procedure will have *some* influence on the decision to operate, and thus on the verification of the diagnosis. But it will not *determine* the decision to operate. The decision to operate is based secondarily on the severity of the complaints and the inclination of the patient to undergo surgery. (Most physicians advise their patient to start with conservative treatment, i.e., bedrest.) Moreover, the radiological diagnosis is not the sole source of information for the physician, and it may be ignored if the symptoms are quite clear.

These three sources of bias also influence each other: for instance, as the selection criteria of the patients (a) become stricter, the non-operative fraction of patients (b) becomes smaller, and the influence of the radiology on the decision to operate (c) decreases.

Randomized study design would overcome many of the afore mentioned problems but, due to ethical and practical problems, this is not generally accepted in this field. Blind experiments are not applicable, but X-ray evaluation without clinical information [7, 11, 13, 38, 39] as well as without influence on the decision to operate [40] have been tried.

The ROC Model

The model underlying traditional ROC analysis has limitations, which are for instance exhaustively described in [25, 41]. For the present type of analysis some limitations are that the (condensed) feature distributions of diseased and normals are not known, and that the disease verfication procedure (surgery) depends upon the diagnostic test and the selection of the patients. The confidence intervals for the resulting odds ratio is not independent of the prevalence, a parameter that is likely to change between samples. A logistic distribution of the condensed features

for normals and diseased was used as an approximation. It is difficult to assess the accuracy of this approximation. A recent paper [32] indicated, however, that the differences between, e.g., binormal, binomial, Poisson or gamma distributions may have little practical consequences.

In a large part of the clinical literature, it is implicitly assumed that a test has certain diagnostic qualities. Social science has taught us to be more aware of the influence of the observer on the results. The ROC analysis expresses the observer only in terms of using a low or high "threshold value" (or cut-off point) for dichotomizing the observations into two classes. Thus, we are both able and bound to consider the diagnostic qualities of a (test, observer) combination. Ideally, the (sensitivity, specificity) vectors can be split into a test related part (the odds ratio component, along the 135 diagonal) and an orthogonal observer-based part (the "threshold value" component, along the 45° diagonal).

Though a logistic distribution of the features seems to be a reasonable approximation, the requirement of an equal variance of normals and diseased is sometimes too strict. For a fully developed and standardized test procedure, the ratio Var(*diseased*)/Var(*normals*) is the square of the slope of a linear regression line fitted to the data. At least for CT this slope is definitely different from 1. This may, however, be due to the fact that the sensitivity of the CT procedure gradually improved. It has to be kept in mind that myelography and phlebography existed for many years before 1970, the period where our review began, while CT only emerged during the latter half of the review period. This hypothesis is, however, not substantiated by a (weighted) logistic regression with our data of the sensitivity against the (approximate) date of investigation mentioned in the various reports (the regression coefficient was close to zero). However, one would expect a large ratio of Var(*diseased*)/Var(*normals*) to be also expressed by other procedures, not just by CT only.

From Fig. 2 the specificity appears clearly lower for CT than for the other procedures. A final hypothesis would be that some other condition might be confused with LDH on CT. Suppose, for example, that the specificity would be uniformly (i.e., independently from the sensitivity) diminished by 10-15% by such an unknown condition. In logistic ROC space 10% or 15% constitutes a larger interval at the edge than in the middle of the scale. Hence, in that case, a straight line with unit slope (running through the left upper part of the diagram) would be transformed into approximately a straight line with a much higher slope, so that this hypothesis is at least consistent with the present data. Clearly, a more detailed investigation would be needed to confirm or reject this hypothesis.

It is remarked that some of the other recent CT reports [38, 39] were not included because of an incomparable presentation of results or a different study design, and that new publications in the next few years on CT and LDH may change the position of CT.

Finally, having placed more emphasis on the test quality and possible observer differences, we only want to mention briefly that there are more aspects to the choice of which radiological procedure to apply. Safety, convenience and costs of the test are among the most important other considerations; these, for example, tend to rate phlebography as being outdated. New reports on CT and the development of new techniques like Magnetic Resonance Imaging may change the order of preference, but currently we conclude that phlebography has the highest discriminatory power, followed by myelography and CT, in visualization of LDH.

Appendix

Probability model and graphical representation

The data of each publication and each diagnostic method from Table 1 are condensed into a 2 × 2 table, see below. Let us first consider the analysis of one such 2 × 2 table. For clearness of presentation, we distinguish between observed values such as $a, b, c, \ldots$ and the associated random variables $A, B, C, \ldots$, and also between parameters, such as $p_a, p_a, p_c, \ldots$ and their estimates $\hat{p}_a, \hat{p}_b, \hat{p}_c, \ldots$. Considering the total number of observations as fixed, we make the usual probabilistic assumption $(A, B, C, D) \sim M(n; p_a, p_b, p_c, p_d)$, which stands for (A, B, C, D) is multinomially distributed with total sample size n and vector of probabilities (p_a, p_b, p_c, p_d). As $p_a + p_b + p_c + p_d = 1$, the model has four independent parameters (n included).

	T+	T−	
D+	a	b	m
D−	c	d	n−m
	a+c	b+d	n

	T+	T−	
D+	p_a	p_b	
D−	p_c	p_d	
			1

Another set of four parameters characterizing the 2 × 2 table is (n, se, sp, pr), where se stands for sensitivity, sp for specificity and pr for prevalence. According to their definition, we have $se = p_a/(p_a + p_b)$, $sp = p_d/(p_c + p_d)$, $pr = p_a + p_b$. The multinomial assumption implies that the marginal distribution of the number of diseased individuals in the sample is binomial, $M \sim B(n, pr)$, and that, conditional on the observed number of diseased, $A \sim B(m, se)$ and $C \sim B(n - m, 1 - sp)$. The usual (frequentist) estimates of (se, sp, pr) are $\widehat{se} = a/m, \widehat{sp} = d/(n - m), \widehat{pr} = m/n$.

In Fig. 2 a graphical view is given of the situation. The axes represent logistic transformations of se and $1-sp$, i.e., $x_2 = \log(se/(1-se))$ is plotted against $x_1 = \log((1-sp)/sp)$. The corresponding unknown parameters of an investigation are denoted by μ_{se} and μ_{sp}. The distributions of the estimators $\hat{\mu}_{se} = \log(\widehat{se}/(1-\widehat{se}))$ and $\hat{\mu}_{sp} = \log((1 - \widehat{sp})/\widehat{sp})$ are aymptotically normal with variances σ^2_{se} and σ^2_{sp}, respectively. The contours of equiprobability are ellipses with half axes proportional to σ_{se} and σ_{sp}. As in practice we condition on the total number of diseased individuals, $\hat{\mu}_{se}$ and $\hat{\mu}_{sp}$ are independent, whence the axes of these ellipses are parallel to the coordinate axes. Now, we have a third set of four parameters characterizing the 2 × 2 table: $(\mu_{se}, \mu_{sp}, \sigma_{se}, \sigma_{sp})$. This set has a clear geometrical

interpretation in ROC space. Expressed in the previous sets of parameters we have

$$\sigma_{se}^2 = n^{-1}(p_a^{-1} + p_b^{-1}) = (n\, pr\, se(1-se))^{-1} \tag{1}$$

and

$$\sigma_{sp}^2 = n^{-1}(p_c^{-1} + p_d^{-1}) = (n(1-pr)sp(1-sp))^{-1}, \tag{2}$$

which is in accordance with the well-known simple estimates $\hat{\sigma}_{se}^2 = a^{-1} + b^{-1}$, and $\hat{\sigma}_{sp}^2 = c^{-1} + d^{-1}$.

Alternatively, one can characterize the size and the shape of an ellipse by the area $V = \pi\sigma_{se}\sigma_{sp}$ and by the ratio $R = \sigma_{se}/\sigma_{sp}$ of the axes, respectively. Obviously,

$$R^2 = (1-pr)sp(1-sp)/pr\, se(1-se) \tag{3}$$

and

$$V^2 = \pi^2/(n^2 pr(1-pr)se(1-se)sp(1-sp)). \tag{4}$$

Hence, the ratio of the axes is independent of the sample size, but the area depends, besides on the sample size, also on (μ_{se}, μ_{sp}) and the prevalence pr. For ellipses having the same center and the same shape R, the area only depends on and is inversely proportional to the sample size.

The center (μ_{se}, μ_{sp}) of an ellipse can also be characterized by

$$(\mu_{OR}, \mu_{TV}) = (\mu_{se} - \mu_{sp}, \mu_{se} + \mu_{sp}), \tag{5}$$

where μ_{OR} stands for the log odds ratio and μ_{TV} for the logarithm of the 'threshold value'. Geometrically this is interpreted as looking at the center from a different coordinate system, in which the OR-axis is constituted by the line $\mu_{se} = -\mu_{sp}$ (i.e., $\mu_{TV} = 0$) and the TV-axis equals the line $\mu_{se} = \mu_{sp}$ (i.e., $\mu_{OR} = 0$). The advantage of this second coordinate system is that $\log OR$ can be viewed as a measure of the performance of the diagnostic procedure, whereas the TV-axis describes the trade-off between *se* and *sp* that occurs whenever the 'observer' changes the (idealized) threshold value or cut-off point of the procedure that differentiates between diseased and non-diseased patients. (When we speak about 'observer' in this section, we actually mean the combination of the observer and the details of the diagnostic test procedure.) The log odds ratio is estimated by $\hat{\mu}_{OR} = \log(ad/bc)$ and σ_{OR}^2 by $\hat{\sigma}_{OR}^2 = a^{-1} + b^{-1} + c^{-1} + d^{-1}$. Similarly, $\hat{\mu}_{TV} = \log(ac/bd)$ and $\hat{\sigma}_{TV}^2 = \hat{\sigma}_{OR}^2$. Obviously, $\hat{\mu}_{OR}$ and $\hat{\mu}_{TV}$ are not independent, their correlation coefficient being $(\sigma_{se}^2 - \sigma_{sp}^2)/(\sigma_{se}^2 + \sigma_{sp}^2)$.

In the practical analysis, for all estimates, 0.5 was added to the four cells in each 2×2 table to avoid problems arising with empty cells. From a frequentist point of view, this generally introduces negligible biases, see, e.g., [42]. From a Bayesian

point of view, it is 'optimal' for a rather conservative choice ($Be(\frac{1}{2}, \frac{1}{2})$) for the prior distribution: a uniform prior would even entail adding +1 to each cell.

Using the normal approximation on logistic scale, we have for k publications and one diagnostic procedure the following general model

$$M_o : X_{1,j} \sim N(\mu_{sp,j}, \sigma^2_{sp,j}), \quad X_{2,j} \sim N(\mu_{se,j}, \sigma^2_{se,j}), \quad j = 1, \ldots, k, \qquad (6)$$

where $(X_{1,j}, X_{2,j})$ denotes the estimator for $(\mu_{sp,j}, \mu_{se,j})$ based on the j^{th} observation. Conditional on the total number of diseased individuals within each study, $X_{1,j}$ and $X_{2,j}$ are independent. We also assume independence of $(X_{1,j}, X_{2,j})$, $j = 1, \ldots, k$ across the publications. It is of interest to consider the sub-model $M_1 : \mu_{se,j} = \alpha + \beta\mu_{sp,j}$ which indicates that in logistic ROC space the pairs $(\mu_{sp,j}, \mu_{se,j})$ lie on a straight line with slope β. Note that this model can be considered as a logistic regression model (for a recent review see [43]), but now with errors in the 'independent' dichotomous variable *sp*. If $\beta = 1$, we have the further restricted sub-model M_2 that the publications show a common log odds ratio α.

In an 'observer' interpretation of model M_1 (all k investigations ideally based on the same patients, whereas the different 'observers' apply different threshold values), $\beta \neq 1$ means that *sp* and *se* are not traded-off equally, but, on logistic scale, with a proportionality factor β. In a 'population' interpretation of model M_1 (ideally the same type of 'observer', the patients of each investigation are a random sample from a large population of diseased and non-diseased individuals) one can, at least under the following idealised, precise specification, give another meaning to β. The two distributions of diagnostic features of the diseased and non-diseased may, of course, be higher dimensional, and could be represented as clouds in IR^p, say. The allocation rules of the different investigations, which decide between $T+$ and $T-$, should be based on the same set of parallel $k-1$ dimensional planes in IR^p, each investigation using its own, fixed plane. By projecting on the line perpendicular to these planes, one gets two univariate feature distributions, which should be approximately logistic. Under these assumptions, the unknown, true $(\mu_{sp,j}, \mu_{se,j})$ values of the k investigations lie on a straight line in logistic ROC-space, with β^2 equal the ratio of the variances of the two projected, univariate feature distributions of the diseased and the non-diseased, respectively.

Estimation

The parameters of model M_1 are estimated by maximum likelihood or, equivalently, as we have normally distributed variables, by generalized least squares.

Denoting for brevity $\mu_{sp,j}$ by $\mu_{1,j}$, $\mu_{se,j}$ by $\mu_{2,j}$, and similarly for σ^2, we have in model M_1

$$-2\log L(\alpha,\beta,\underline{\mu}) = SS(\alpha,\beta,\underline{\mu}) = \sum_{j=1}^{k}((\frac{(X_{1,j}-\mu_{1,j})}{\sigma_{1,j}})^2 + (\frac{(X_{2,j}-\alpha-\beta\mu_{1,j}}{\sigma_{2,j}})^2), \tag{7}$$

where L stands for the likelihood, SS for the residual sum of squares and $\underline{\mu} = (\mu_{1,1},\ldots,\mu_{1,k})$ for the vector of the k logistically transformed sensitivities. The minimum of this expression determines the maximum likelihood (ML) estimates for α, β and $\underline{\mu}$. (For simplicity we suppose the parameters $\sigma_{1,j}$ and $\sigma_{2,j}$ to be known, inserting at a later stage the usual estimates for these parameters under model M_o.)

We first minimize with respect to $\underline{\mu}$ for fixed values of α and β. Solving $\partial/\partial\mu_{1,j}SS(\alpha,\beta,\underline{\mu}) = 0$ gives, after some calculation,

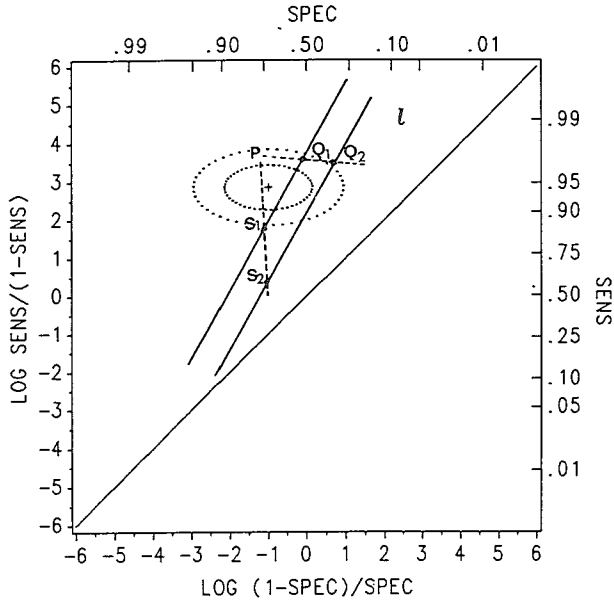
$$\hat{\mu}_{1,j} = \frac{X_{1,j}\sigma_{2,j}^2 + (X_{2,j}-\alpha)\beta\sigma_{1,j}^2}{\sigma_{2,j}^2 + \beta^2\sigma_{1,j}^2}. \tag{8}$$

Inserting these estimates into (7), we get the following 'likelihood profile'

$$-2\log L(\alpha,\beta,\hat{\underline{\mu}}) = SS(\alpha,\beta,\hat{\underline{\mu}}) = \sum_j \frac{(X_{2,j}-\alpha-\beta X_{1,j})^2}{\sigma_{2,j}^2 + \beta^2\sigma_{1,j}^2}. \tag{9}$$

This expression can directly be interpreted as the sum of squares of (generalised) distances between the observations $(X_{1,j},X_{2,j})$ and the line $x_2 = \alpha + \beta x_1$. For the j^{th} observation this distance is defined by the metric induced by the ellipses of equiprobability, whose half axes are proportional to $\sigma_{1,j}$ and $\sigma_{2,j}$. All points (x_1,x_2) on the elliptical contour $(\frac{x_1-X_{1,j}}{\sigma_{1,j}})^2 + (\frac{x_2-X_{2,j}}{\sigma_{2,j}})^2 = 1$ have the same, say unit, distance to $(X_{1,j},X_{2,j})$. The generalized distance between $(X_{1,j},X_{2,j})$ and the line $x_2 = \alpha + \beta x_1$ is the factor with which the unit ellipse has to be blown-up to touch the line. From Fig. 3 one can see that this distance equals $PQ_2/PQ_1 = PS_2/PS_1$. PS_2 is equal to (the absolute value of) $X_{2,j}-(\alpha+\beta X_{1,j})$. To calculate PS_1 we choose for convenience a new coordinate system with origin in $(X_{1,j},X_{2,j})$. As we know that the tangent lines to the ellipse $\frac{x_1'^2}{\sigma_1^2} + \frac{x_2'^2}{\sigma_2^2} = 1$ with slope β are given by $x_2' = \beta x_1' \pm (\beta^2\sigma_1^2 + \sigma_2^2)^{1/2}$, we have (take $x_1' = 0$)

Figure 3 – Generalized distance between observation P and regression line *l*.



$PS_1 = (\beta^2\sigma_1^2 + \sigma_2^2)^{1/2}$, hence $PS_2/PS_1 = |(X_{2,j} - (\alpha + \beta X_{1,j})|/(\sigma_{2,j}^2 + \beta^2\sigma_{1,j}^2)^{1/2}$ and the sum of (generalized) squared distances is given by (9).

For a fixed value of β, one can minimize (9) with respect to α. Solving $\partial/\partial\alpha SS(\alpha, \beta) = 0$ leads to the expression

$$\hat{\alpha}(\beta) = \sum_j \frac{X_{2,j} - \beta X_{1,j}}{\sigma_{2,j}^2 + \beta^2\sigma_{1,j}^2} \Big/ \sum_j \frac{1}{\sigma_{2,j}^2 + \beta^2\sigma_{1,j}^2}. \tag{10}$$

Note that $\alpha(1)$, being equal to the average of the individual log odds ratios weighted inversely proportional to their variances, is just the 'logit estimator' for the common log odds ratio, implemented in SAS [44]. Obviously, $\alpha(0)$ is the estimator of a common μ_{se} and for $\beta \to \infty$, $\hat{\alpha}(\beta)/\beta \to -\mu_{sp}$.

Inserting (10) into (9), we finally get a sum-of-squares profile $SS(\hat{\alpha}(\beta), \beta)$ for β alone. Unfortunately, no closed solution exists for $\partial/\partial\beta\, SS(\hat{\alpha}(\beta), \beta) = 0$, whence the sum-of-squares profile has to be minimized numerically.

Figure 4 - Polar likelihood contour for the slope of the regression line; myelography.

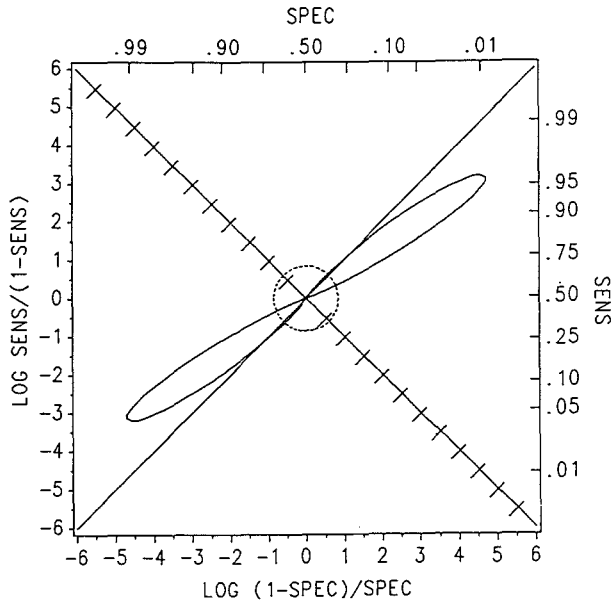
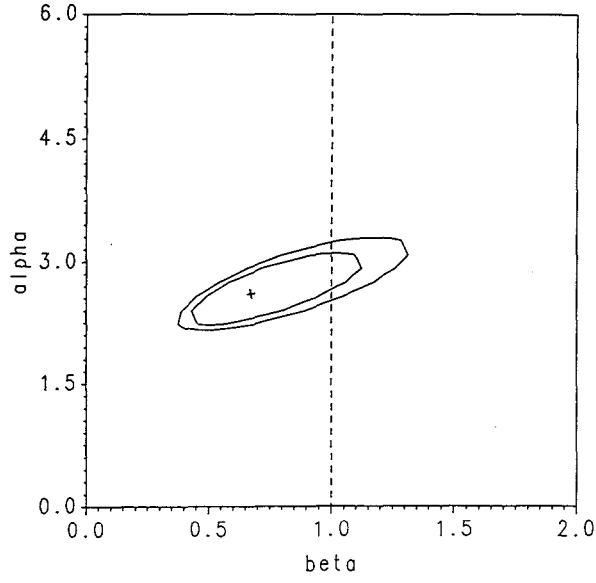


Figure 5 – Residual sum of squares as a function of the intercept α and slope β of the regression line for myelography in logistic ROC space.



Testing and confidence intervals

Once the global minimum of (7) is found, the adequacy of model M_1 (indicating that the points $(\mu_{se,j}, \mu_{sp,j}), j = 1, \ldots, k$ are connected by an arbitrary straight line) can be tested by noting that if M_1 holds, $SS(\hat{\alpha}, \hat{\beta}, \hat{\underline{\mu}})$ is distributed as $\chi^2_{2n-(n+2)} = \chi^2_{n-2}$. For $\beta = 1$, this test, now with $n - 1$ df, corresponds asymptotically to the Breslow-Day test for the homogeneity of odds-ratios [35], implemented in SAS [44].

According to the likelihood-ratio principle, the null-hypothesis $H_o : \beta = \beta_o$ is not rejected, at asymptotically the 5% level, as long as

$$-2 \log L(\beta_o)/L(\hat{\beta}) = SS(\beta_o) - SS(\hat{\beta}) \leq \chi^2_{1;.05} = 3.84. \qquad (11)$$

In this way, an approximate 95% confidence interval for β is formed. It is sometimes convenient to use the angle ϕ, where $\tan \phi = \beta$, instead of β: an interval containing $\beta = \pm\infty$ is transformed to a regular, compact interval for ϕ. To give a pictorial representation, one can make a polar plot of the likelihood $L(\phi)$ against ϕ, normalized, e.g., by $L(\phi) = 1$ (see Fig. 4). This gives a visual impression in how far various slopes of the regression line are compatible with the available data, and one can, more formally, read off (aymptotic) confidence intervals for the 'true' slope.

To find a confidence interval for α, one could, in a similar way, plot the profile $SS(\alpha, \hat{\beta}(\alpha))$ as a function of α. Now, however, $\hat{\beta}(\alpha)$ has to be computed numerically for each α and then inserted into $SS(\alpha, \beta)$.

Alternatives

An alternative method which provides directly estimates and confidence intervals for α and β, consists in plotting $SS(\alpha, \beta)$ as a function of α and β (see Fig. 5). In this approach, much work is taken over by the contour plotting program. The ML estimates α and β are determined by the position of the (global) minimum, and an approximate 95% confidence interval for α and β is read off by projecting the contour $SS = SS_{min} + 3.84$ on the coordinate axes. Finally, by the likelihood ratio principle, an approximate 95% *simultaneous* confidence interval for α and β is formed by the region inside the contour $SS = SS_{min} + \chi^2_{2;.05} = SS_{min} + 5.9$. Similarly, SS can be plotted as a function of $\delta = |\alpha| / \sqrt{1 + \beta^2}$ and $\phi = \arctan \beta$, where δ represents the perpendicular distance from the regression line to the origin. This distance is one measure of the discriminatory power of the procedure, to be preferred over the log odds ratio if the regression slope deviates from unity. The area under the ROC curve in the standard ROC space can, for each fixed ϕ, be written as a monotonic function of δ. This function, which can be computed

numerically, turns out to be almost independent of ϕ. Hence, comparing diagnostic procedures in logistic ROC space on the basis of δ is to a good approximation equivalent to comparing them by the area under the ROC curve. As can easily be seen, in the normally transformed ROC space the area under the ROC curve is completely independent of ϕ.

References

[1] Van Romunde LKJ, Rouwens, T. Trends in frequentie van ziekten in Nederland. In: Grobbee DE, Hofman A, eds. *Epidemiologie van ziekten in Nederland.* Utrecht: Wetenschappelijke Uitgeverij Bunge, 1989: 1-16.

[2] Weber H. Lumbar disc herniation — a controlled, prospective study with ten years of observation. *Spine* 1983; **8**: 131-140.

[3] Hampton AO, Robinson JM. Roentgenographic demonstration of rupture of the intravertebral disk into the spinal canal after injection of Lipiodol. *AJR* 1936; **36**: 782-803.

[4] Kardaun JWPF, Schipper J, Braakman R. CT, myelography, and phlebography in the Detection of lumbar disk herniation: An analysis of the literature. *Am J Neuroradiol* 1989; **10**: 1111-1122.

[5] Einarson TR, McGhan WF, Bootman JL, Sabers DL. Meta-analysis: quantitative integration of independent research results. *Am J Hosp Pharm* 1985; **42**: 1957-64.

[6] Jenicek M. Meta-analysis in medicine — where we are and where we want to go. *J Clin Epidemiol* 1989; **42**: 35-44.

[7] Lotz PR, Seeger JF, Gabrielsen TO. Prospective comparison of epidural venography and iophendylate myelography in the diagnosis of herniated lumbar disks. *Radiology* 1980; **134**: 127-32.

[8] Schipper J, Kardaun JWPF, Braakman R, Van Dongen KJ, Blaauw G. Lumbar disk herniation: Diagnosis with CT or Myelography? *Radiology* 1987; **165**: 227-231.

[9] Hudgins WR. The predictive value of myelography in the diagnosis of ruptured lumbar discs. *J Neurosurgery* 1970; **32**: 152-62.

[10] Drasin GF, Daffner RH, Sexton RF, et al. Epidural venography: diagnosis of herniated lumbar intervertebral disc and other disease of the epidural space. *AJR* 1976; **126**: 1010-16.

[11] MacNab I, St. Louis EL, Grabias SL, et al. Selective ascending lumbosacral venography in the assessment of lumbar-disc herniation. *J Bone Joint Surg* 1976; **58A**: 1093-8.

[12] Miller MH, Handel SF, Coan JD. Transfemoral lumbar epidural venography. *AJR* 1978; **126**: 1003-9.

[13] Mohsenipour I, Twerdy K, Pirker E. Eine Gegenüberstellung Röntgenbild - Operationsbefund bei lumbalen Bandscheibenerkrankungen. *RöFo* 1977; **127**: 540-3.

[14] Moringlane JR, Voigt K, Seeger W. Vergleich myelographischer und intraoperativer Befunde beim lumbalen Bandscheibenvorfall. *Neurochirurgia* 1977; **20**: 199-208.
[15] Cook PL, Wise K. A correlation of the surgical and radiculographic findings in lumbar disc herniation. *Clin Radiol* 1979; **30**: 671-82.
[16] Meyenhorst GCH. *Transfemoral epidural double-catheter venography in the diagnosis of lumbar disc herniation.* Thesis, Nijmegen Catholic University, Netherlands, 1979.
[17] Thijssen HOM, Rombouts JJM, Walder HAD. Diagnostic accuracy of lumbar myelography in the detection of lumbar disk herniations. *Diagn Imaging* 1980; **49**: 188-92.
[18] Claussen C, Grumme T, Treisch J, Lochner B, Katzner E. Die Diagnostik des lumbalen Bandscheibenvorfalls. *RöFo* 1982; **136**: 1-8.
[19] Fries JW, Abodeely DA, Vijungco JG, et al. Computed tomography of herniated and extruded nucleus pulposus. *J Comput Assist Tomogr* 1982; **6**: 874-887.
[20] Haughton VM, Eldevik OP, Magnaes B, et al. A prospective comparison of computed tomography and myelography in the diagnosis of herniated lumbar disks. *Radiology* 1982; **142**: 103-10.
[21] Jepson K, Nada A, Rymaszewski L. The role of radiculography in the management of lesions of the lumbar disc. *J Bone Joint Surgery* 1982; **64B**: 405-8.
[22] Moufarrij NA, Hardy Jr RW, Weinstein MA. Computed tomographic, myelographic and operative findings in patients with suspected herniated lumbar discs. *Neurosurgery* 1983; **12**: 184-8.
[23] Valat JP, Gatti P, Saindelle A. Phlébographie lombaire. Apport au diagnostic des lombosciatiques d'origine discale. *Sem Hôp Paris* 1984; **60**: 539-42.
[24] Kampmann H, Schroedl P, Spranger M. Diagnostik lumbaler Bandscheibenvorfälle durch Computertomographie. Eine klinische Vergleichsstudie zwischen myelographischen, computertomographischen und operativen Untersuchungsergebnissen von 158 Patienten. *Röntgenblätter* 1985; **38**: 387-91.
[25] Swets JA. Indices of discrimination or diagnostic accuracy: ROC and implied models. *Psych Bull* 1986; **99**: 100-107.
[26] Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; **8**: 283-98.
[27] Lusted LB. *Introduction to medical decision making.* Springfield, IL: CC Thomas Pub., 1968.
[28] *Wissenschaftliche Tabellen Geigy, Teilband Statistik,* 8. Aufl. Basel: Ciba Geigy, 1980.
[29] Koopman PAR. Sensitiviteit, specificiteit, aannemelijkheidsverhouding en voorspellende waarde van diagnostische testen. *Kwantitatieve Methoden* 1983; **9**: 39-57.
[30] Green DM, Swets JA. *Signal detection theory and psychophysics.* Huntington, NY: Robert E. Kriegler Pub. Co., 1974.
[31] Johnson NL, Kotz S. *Distributions in statistics; continuous univariate distributions 2.* New York: Wiley & Sons, 1970.

[32] Hanley JA. The robustness of the "binormal" assumptions used in fitted ROC curves. *Med Decis Making* 1988; **8**: 197-203.
[33] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719-48.
[34] Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research; Principles and quantitative methods.* Belmont, CA: Wadsworth, Inc, 1982.
[35] Breslow NE, Day NE. *Statistical methods in cancer research, volume 1: The analysis of case-control studies.* Lyon: IARC, 1980.
[36] Begg CB, Greenes RA. Biases in the assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 1983; **39**: 207-15.
[37] Howe J, Frymoyer JW. The effects of questionnaire design on the determination of end results in lumbar spinal surgery. *Spine* 1985; **10**: 304-5.
[38] Bell GR, Rothman RH, Booth RE, et al. A study of computer assisted tomography in the diagnosis of herniated disk and spinal stenosis. *Spine* 1984; **9**: 552-6.
[39] Modic MT, Masaryk T, Boumphrey F, Goormastic M, Bell G. Lumbar herniated disk disease and canal stenosis: prospective evaluation by surface coil MR, CT, and myelography. *AJR* 1986; **147**: 757-65.
[40] Esperen JO, Kosteljanetz M, Halaburt H, Miletic T. Predictive Value of radiculography in Patients with lumbago-sciatica. A prospective study. *Acta Neurochir* 1984; **73**: 213-21.
[41] Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987; **6**: 411-23.
[42] Bedrick EJ. Estimating the variance of empirical logits and contrasts in empirical log probabilities. *Biometrics* 1984; **40**: 805-9.
[43] Houwelingen JC van, Cessie S le. Logistic regression; a review. *Statistica Neerlandica* 1988; **42**: 215-32.
[44] *SAS User's Guide; Statistics.* Cary NC: SAS Institute Inc., 1985.

4

CHEMONUCLEOLYSE WEGENS HERNIA NUCLEI PULPOSI LUMBALIS; WERKZAAMHEID EN BIJWERKINGEN.

JWPF Kardaun [1]
J Schipper [2]

[1] Instituut Maatschappelijke Gezondheidszorg,
Erasmus Universiteit, Rotterdam

[2] Afdeling Radiodiagnostiek,
Academisch Ziekenhuis, Leiden.

Summary

Chemonucleolysis, a treatment for intervertebral disk herniation by injection of a proteolytic enzyme in the herniated disk, is a possible alternative for surgical treatment. The qualities of chemonucleolysis are evaluated on the aspects of effectiveness and complications, based upon the large amount of available literature. The effectiveness has been compared in randomized clinical trials with a placebo and with surgery. Chemonucleolysis appeared to have a clearly and significantly higher probability of a successful result one year after treatment, compared with placebo, but a lower probability compared with surgery. A larger number of non-randomized and often non-comparative studies of chemonucleolysis exists, that give a remarkably similar distribution of the success-rate, whether split in early (pre-1985) or late, and in short term (less than 2 year) or long term follow-up. The median success-rate is approximately 75%. This could be a characteristic of the treatment, but may also be an indication that the method of evaluation is not very sensitive.

The complications are dominated by an overall mortality of approximately 0.5 per thousand, and the possibility of an anaphylactic reaction to chymopapain, which has been estimated at 4—7 per thousand. Recent developments, including a skin test on specific IgE and the use of H_1-H_2 antagonists, may reduce the frequency of this complication.

(This summary was not included in the text published in the Nederlands Tijdschrift voor Geneeskunde.)

Published in the Nederlands Tijdschrift voor Geneeskunde 1987; **132**: 285-289.

Inleiding

De meest gebruikelijke behandelingen van de hernia nuclei pulposi zijn de conservatieve therapie (vooral bedrust, vaak gevolgd door fysiotherapie) en de operatieve behandeling (verwijderen van de aangetaste discus). Daarnaast is sinds 1964 een behandeling mogelijk waarbij chymopapaïne, een eiwitsplitsend enzym, in de aangetaste discus wordt ingespoten.[1] Hoewel het werkingsmechanisme niet precies vaststaat, is het zeer aannemelijk dat injectie van dit enzym door lysis van de glycoproteïnen een vermindering van de turgor van de discus en zo van de herniatie geeft.[2] Deze behandeling, chemonucleolyse, heeft in de afgelopen twee decennia een levendige discussie gaande gehouden, onder meer omdat de effectiviteit in twijfel werd getrokken[3-5] en omdat er nogal wat bijwerkingen werden gemeld, waaronder anafylactische reacties.[6,7] Om deze reden is het middel enige tijd (1975-1982) in de V.S. verboden geweest, maar na een tweetal gerandomiseerde klinische onderzoeken werd het vrijgegeven.[8,9] In Nederland is het middel sinds juni 1986 geregistreerd.

De vraag naar het toepassingsgebied van deze methode is daarmee nog niet opgelost, zodat er behoefte bestaat de indicatiestelling, de werkzaamheid en de complicaties van deze behandeling opnieuw te bezien. In dit tijdschrift is aangedrongen op een strikte indicatiestelling, dat wil zeggen alleen bij die patiënten waarbij een discushernia is aangetoond, met uitsluiting van andere aandoeningen,[10,11] en zelfs alleen bij een deel van deze patiënten.[12]

Dit overzicht spitst zich toe op de effecten en de bijwerkingen van chemonucleolyse. Het is een uitvloeisel van een adviesaanvraag aan de Gezondheidsraad om voorwaarden te stellen aan de toepassing van deze behandelingsvorm.[13] Daartoe werd de beschikbare literatuur geanalyseerd. Er wordt geen aandacht besteed aan de werkwijze van het middel, de uitvoering of de formulering van een indicatiestelling.

Vooral het laatste is van belang voor een optimaal resultaat; de literatuur geeft daarvoor echter geen goede aanknopingspunten. De vraag is of uit de veelheid aan literatuur — in 1986 verschenen meer dan 48 artikelen over dit onderwerp — een duidelijk beeld over de effectiviteit naar voren komt, welke bijwerkingen te verwachten zijn en hoe deze laatste te vermijden zijn.

Effectiviteit

Gerandomiseerde studies

Bij aandoeningen met langdurige en wisselende pijn, zoals het radiculaire compressiesyndroom, neemt de placebocomponent van een behandeling een belang-

rijke plaats in. De werkzaamheid van chymopapaïne is dan ook het beste vast te stellen door middel van een dubbelblind gerandomiseerd onderzoek. Patiënten die niet in de chymopapaïnegroep geloot worden, krijgen een intradiscale injectie met een zoutoplossing,[8,9] of een ander, in principe werkzaam, preparaat (hydrocortison[14]).

In vier onderzoekingen (tabel 1) met dubbelblinde, gerandomiseerde opzet, bleek chemonucleolyse telkens effectiever dan een placebo, hoewel het verschil in de eerste studie niet groot was.[3,8,9,15] Uit de gezamelijke studies blijkt, dat de behandeling met chymopapaïne de kansverhouding (*odds*) voor een succesvol resultaat met een factor 2,6 (1,6 — 4,3) verbetert t.o.v. een behandeling met een placebo. Hoewel dit een duidelijk verschil is, moeten hierbij twee kanttekeningen gemaakt worden. Enerzijds heeft de behandeling met alleen een placebo opmerkelijk veel "effect" (42% — 60%), anderzijds betreft het alleen een succes op de middellange termijn (tot 1 jaar).

Een behandeling moet niet alleen de vergelijking met een placebo doorstaan, maar ook die met de andere in aanmerking komende therapieën. Als gerandomiseerd wordt tussen chemonucleolyse en operatie is een dubbelblind onderzoek niet uitvoerbaar, zodat de uitkomsten mogelijk beïnvloed zijn door verwachtingen van arts en patiënt aangaande de behandeling.[4,5] Deze onderzoeken melden beide een duidelijk effectievere werking van operatie. Bij Crawshaw et al. waren de resultaten: bij 44% succes met chemonucleolyse, in 85% met operatie.[5] Bij Ejeskær et al. waren de resultaten van chemonucleolyse zo slecht dat dit een reden was het onderzoek af te gelasten.[4] Hoewel deze resultaten duidelijk zijn, betreft het slechts 80 patiënten (zie tabel 1).

Niet-gerandomiseerde studies

Bij de vele studies die niet gerandomiseerd zijn, treden twee problemen op die een vergelijking bemoeilijken, nl. de selectie van de patiënten vóór de behandeling en het bepalen van een succesvol resultaat. Een ruime indicatie voor chemonucleolyse omvat "low back pain" en asymptomatische vormen van hernia.[16] Strenge varianten eisen klinische en radiologische aanwijzingen voor een hernia nuclei pulposi, gepaard met uitsluiting van arbeidsongeschiktheidskwesties of met positieve discografie en een vergeefse bedrustkuur.[18,19]

De meeste onderzoekingen nemen als criterium voor succes de subjectieve tevredenheid van de patiënt in 3 tot 5 graden, enige maanden tot 10 jaar na de behandeling. Het meten van de tevredenheid van de patiënt met behulp van meer objectieve criteria, zoals gespecificeerde pijnklachten, werkhervatting, neurologische symptomen, enz. geeft geen duidelijker beeld.[8,9,16] De klachten bij een radiculair compressiesyndroom zijn niet eens zo menigvuldig, maar het patroon is blijkbaar toch te complex om de klachten en de verbeteringen ervan tot een schaal

voor succes van de behandeling te kunnen herleiden. De tevredenheid van de patiënt als ijkmaat heeft ook nogal wat bezwaren, omdat tevredenheid niet altijd van een verbetering afhangt. Bij Smyth et al. bijv. wordt de vraag naar het resultaat van de chemonucleolyse door 43% van de patiënten beantwoord met: compleet genezen, terwijl toch 83% van dezelfde patiënten nog last van "any recurrence/persistence of back or leg pain" zegt te hebben.[20]

Volgens Howe en Frymoyer is het dan ook mogelijk om een zeer groot gedeelte van de verschillen uit de onderzoekingen naar de resultaten van een discusherniabehandeling te verklaren uit de keuze van de beoordelingsvragen.[21] Zij verrichtten een follow-up onderzoek van meer dan 10 jaar bij 207 patiënten, die zij willekeurig beoordelingsvragen uit 14 andere onderzoeken voorlegden. De succespercentages liepen uiteen van 60—97; het beste resultaat gaf de subjectieve

Tabel 1. Vergelijking van de effectiviteit van chemonucleolyse met andere therapieën in de literatuur

literauur	jaar van publikatie	aantal patiënten	vergelijking chemonucleolyse met	succes percentage
gerandomiseerde vergelijking met placebo				
Martins et al.[3]	1978	66	placebo	58 - 49
Javid et al.[8]	1983	108	placebo	73 - 42
Fraser[9]	1984	60	placebo	80 - 60 *
Feldman et al.[15]	1986	39	placebo	65 - 42
gerandomiseerde vergelijking met een andere therapie				
Crawshaw et al.[5]	1984	52	operatie	44 - 85
Ejeskær et al.[4]	1983	29	operatie	**
Graham[14]	1976	40	hydrocortison	***
niet gerandomiseerde vergelijking met een andere therapie				
Weinstein et al.[16]	1986	85 en 71	operatie	†
Maroon en Abla [17]	1985	50 en 60	micro-operatie	58 - 90

* In deze studie worden verschillende succespercentages gegeven; de beste waarden zijn vermeld. De laagste combinatie van succespercentages is respectievelijk 63 en 27.

** de resultaten van chemonucleolyse waren zo slecht, dat het onderzoek voortijdig werd afgebroken.

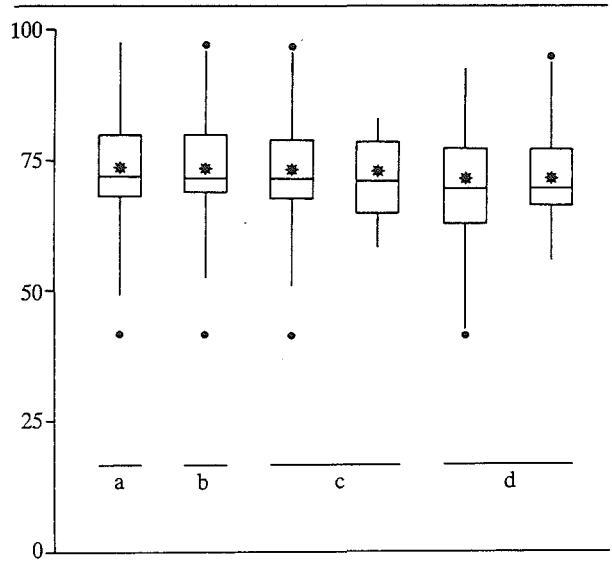
*** de twee therapieën waren gelijkwaardig; succespercentages werden niet gegeven.

† Vele succespercentages van detailmetingen zijn vermeld, die echter niet tot een gezamelijk succespercentage herleid kunnen worden.

tevredenheid van de patiënt te zien, het slechtste de objectief gemeten verbetering. Hieruit concluderen dat de resultaten niet gemeten kunnen worden gaat echter wat ver, en gaat voorbij aan de mogelijkheid om door het grote aantal beoordelingen van chemonucleolyse in de literatuur toch een aardige indruk te krijgen van de te verwachten resultaten. Van de meer dan 40 studies kan verwacht worden, dat zowel de reëel te verwachten variatie aan strenge en milde succescriteria als ook aan stricte en ruime indicatiestellingen vertegenwoordigd is.[22]

De succespercentages van studies die de effectiviteit van chemonucleolyse zonder randomisatie of vergelijkende therapie bepalen, variëren van 98 tot 44 (zie tabel 1). Dit zijn echter extreme waarden: 80% van de resultaten ligt tussen de 60 en 90%. De helft van de resultaten ligt tussen de 70 en 81%; het gemiddelde is 73 procent (figuur, a). Als de studies met grote aantallen zwaarder worden meegeteld (figuur, b) verandert dit niet duidelijk. Om na te gaan of een gedeelte van deze nogal grote variabiliteit verklaard zou kunnen worden door verbeteringen in bijv.

Succespercentages van studies over uitsluitend chemonucleolyse patiënten. De percentages van 41 studies zijn viermaal uitgezet: in (a) alle studies, in (b) eveneens alle studies, waarbij die met grotere patiëntenaantallen zwaarder wegen (evenredig met de wortel uit het aantal patiënten), in (c) gesplitst naar recente (n=13) en vroege (n=28) artikelen (voor of na ultimo 1984), in (d) gesplitst naar lange of korte follow-up (grens: 2 jaar). De zgn box and whisker plots moeten als volgt gelezen worden: de * in de figuren is het gemiddelde; in de rechthoek valt de centrale helft van alle waarnemingen; de lijn in de rechthoek is de mediaan; de uitlopers geven het bereik weer; als de uiterste waarden erg ver weg liggen worden ze met een afzonderlijke • getekend (naar Kardaun)[22].



de kwaliteit van de behandeling of een betere indicatiestelling, is een onderscheid gemaakt tussen de recente en vroege literatuur (grens: eind 1984). Men kan nl. aannemen dat recente behandelingen van de ervaringen op grond van eerdere behandelingen hebben kunnen profiteren, en dat door deze indeling de verbetering zichtbaar zou moeten worden. Deze onderverdeling laat echter nauwelijks verschil zien (figuur, c), evenmin als een opsplitsing van de literatuur in een groep met kort en een met lang follow-up onderzoek (grens: 2 jaar; figuur, d). Ook als rekening wordt gehouden met het feit dat recente literatuur met een lang follow-up onderzoek eigenlijk over "vroege" behandelingen gaat, komt er nauwelijks een verschil naar voren. Er blijkt bovendien dat slechts enkele recente publikaties een kort follow-up onderzoek beschrijven. Dat het onderscheid tussen recente en vroege literatuur en tussen follow-up onderzoek op lange en op korte termijn zo weinig van de verschillen verklaart, ondersteunt de veronderstelling dat er andere zaken zijn, zoals de selectie- en de beoordelingscriteria, die van meer invloed zijn. Voor het bepalen van een verschil tussen resultaten op korte en op lange termijn, kunnen we teruggrijpen naar onderzoeken, gering in aantal, waarin dezelfde patiënten na een kort en na een lang follow-up onderzoek op dezelfde wijze beoordeeld zijn. Deze onderzoeken geven deels uiteenlopende conclusies. Weinstein et al. beschrijven dat het resultaat na enige maanden optimaal is en daarna afneemt; in hun studie heeft 30% van de patiënten minder dan drie jaar baat bij de behandeling.[16] Maciunas en Onofrio vermelden vrij constante resultaten voor een follow-up onderzoek van 1, 5 en 10 jaar voor klachten en beperkingen in het werk.[23] Ook Javid geeft zeer constante waarden voor een follow-up onderzoek van 1 jaar, 3 — 6 jaar en 9 — 12 jaar.[24,25] Jabaay geeft echter betere resultaten bij een follow-up onderzoek van 8 — 10 jaar dan na 1 — 2 jaar.[26]

Een definitieve conclusie over het verloop van de resultaten is dus moeilijk. Toch is dit een belangrijk punt, omdat van de klachten bij een radiculair compressiesyndroom het wisselende en terugkerende karakter welbekend is. Een adaptatiemechanisme van patiënten aan de meeste chronische ziekten zal ook bij de radiculaire compressie een rol spelen.

Bijwerkingen

Na incidentele meldingen van bijwerkingen,[27,28] inventariseerde Watts in 1977 de bijwerkingen systematisch bij 13.700 patiënten (tabel 2).[6] Hij kwam tot een zeer negatief oordeel over chemonucleolyse, wegens een groot scala aan ernstige bijwerkingen die niet gerechtvaardigd zouden zijn door de geringe effectiviteit, zoals die op dat moment net door een gerandomiseerde clinical trial was gemeten.[3] Naderhand zijn door middel van een speciaal "post-marketing surveillance"-programma van een van de fabrikanten de bijwerkingen voor een nog groter aantal

Tabel 2. Overzicht van complicaties van chemonucleolyse.

complicatie	uit Watts[6]		uit Agre at al.[7]	
	per 1000	aantal	per 1000	aantal
overlijden (totaal)	0,58	8	0,38	11
niet t.g.v. shock en bloeding	0,37	5	0,24	7
anafylactische shock	4,0	55*	6,7	194
w.o. sterfgevallen	0,15	2	0,07	2
paraplegieën	0,07	1	0,45	13
cerebrale bloeding	0,44	6	0,21	6
w.o. sterfgevallen	—	—	0,10	3
overige ernstige neurologische aandoeningen	2,1	29	0,14	4
discitis, al of niet steriel	1,6	22	0,76	22

*Dit cijfer is voor ernstige overgevoeligheidsreacties. Als alle reacties worden geteld, is het aantal 207, ofwel 15,2 per 1000.

behandelingen geanalyseerd (zie tabel 2).[7] Nader onderzoek concentreerde zich vooral op het mechanisme en de preventie van de anafylactische reakties.[29-31] Daarbij werd de invloed van factoren als vorm van anesthesie (algeheel of locaal),[7,29,32] H_1- of H_2-receptorblokkerende middelen in de premedicatie,[29,32] en onderzoek naar specifieke IgE-allergie bepaald.[30,31,33,34]

Volgens Moss et al. nam het aantal anafylactische reakties in de periode na 1982 met ongeveer de helft af, waarschijnlijk door het meer toegepast worden van H_1- en H_2-receptorblokkerende middelen en tests op allergie d.m.v. een huidtest of in vitro tests.[29] De recente gegevens over grote aantallen patiënten (meer dan 35.000) maken aannemelijk dat deze tests inderdaad nuttig zijn en het aantal anafylactische reakties met ongeveer de helft kunnen verminderen, terwijl slechts 1% van de patiënten ten onrechte niet voor een chemonucleolyse in aanmerking komt (fout-positieve testuitslagen).[31,33] Bij het ontstaan van complicaties speelt onervarenheid van behandelaars met chemonucleolyse een rol.[34,35] Dit geeft aan dat het aanleren van deze behandelingsmethode bij voorkeur dient te geschieden in instituten waar deze ervaring al aanwezig is.

De kans op complicaties na chemonucleolyse is op zich geen reden deze behandeling af te wijzen. Ze dient gerelateerd te worden aan de ernst van de klachten waarvoor hulp gevraagd wordt, aan de effectiviteit van de behandeling, maar ook aan de kans op complicaties en de effectiviteit van de andere meest in aanmerking komende therapie, operatie. Voor betrekkelijk lichte operaties als de

discusuitruiming zijn de bijwerkingen niet zo degelijk geïnventariseerd als voor chemonucleolyse. Wellicht dat voor chymopapaïne de aansprakelijkheid van de fabrikant, die vooral in de V.S. een grote rol speelt, als positief bij-effect heeft gehad dat de bijwerkingen systematisch geïnventariseerd werden. Zou dit ook voor de operatieve behandeling gebeuren, dan zou de hoeveelheid complicaties waarschijnlijk veel hoger uitvallen dan nu impliciet wordt aangenomen. (Een overzicht van de wat oudere literatuur over de operatieve complicaties geeft Spangfort.)[36] Voordat een beslissing over het toepassingsgebied van chemonucleolyse t.o.v. operatie kan worden afgerond, zullen de behandelingen op ook dit punt vergeleken moeten worden.

Er is veel literatuur beschikbaar voor het bestuderen van de effectiviteit en bijwerkingen van chemonucleolyse. Deze methode is effectiever dan placebobehandeling en minder effectief dan operatieve behandeling. De resultaten van een injectiebehandeling met een placebo zijn echter ook opvallend goed. Een duidelijk verloop van de behandelingsresultaten in de loop van de jaren na chemonucleolyse komt niet tot uiting. De resultaten van 40 studies geven een ruime spreiding van de uitkomsten, waarbij een succespercentage van 70 tot 81 in het belangrijkste deel van de studies wordt gevonden. Het variabele patroon van klachten en van vermindering van klachten en de bijbehorende moeilijkheidsgraad om het resultaat vergelijkbaar te meten, kan hiervoor heel goed een verklaring zijn.

De bijwerkingen van chemonucleolyse zijn goed geïnventariseerd. De belangrijkste zijn een bruto sterfte van 0,5‰, met diverse oorzaken, en een anafylactische reactie (tenminste 0,5%). Daarnaast bestaat nog een groot scala van zeldzamere complicaties. Het gebruik van moderne tests op specifiek IgE lijkt het vóórkomen van anafylactische reakties te kunnen halveren.

Literatuur

1 Smith L: Enzyme dissolution of the nucleus pulposus. *JAMA* 1964; **187**: 137-140.
2 Bradford DS, Oegema TR, Cooper KM, Wakano K, Chao EY: Chymopapain, chemonucleolysis and nucleus pulposus degeneration. A biochemical and biomechanical study. *Spine* 1984; **9**: 135-147.
3 Martins AN, Ramirez A, Johnston J, Schwetschenau PR: Double blind evaluation of chemonucleolysis for herniated lumbar discs: late results. *J Neurosurg* 1978; **49**: 816-827.
4 Ejeskær A, Nachemson A, Herberts P, Lysell E, Andersson G, Irstam L: Surgery versus chemonucleolysis for herniated lumbar discs. A prospective study with random assignment. *Clin Orthop* 1983; **174**: 236-242.

5 Crawshaw C, Frazer AM, Merriam WF, Mulholland RC, Webb JK: A comparison of surgery and chemonucleolysis in the treatment of sciatica. *Spine* 1984; **9**: 195-198.

6 Watts C: Complications of chemonucleolysis for lumbar disc disease. *Neurossurgery* 1977; **1**: 2-5.

7 Agre K, Wilson RR, Brim M, McDermott DJ: Chymodiactin postmarketing surveillance — Demographic and adverse experience data in 29,075 patients. *Spine* 1984; **9**: 479-485.

8 Javid MJ, Nordby EJ, Ford LT, Hejna WJ, Whisler WW, Burton C, et al.: Safety and efficacy of chymopapain (chymodiactin) in herniated nucleus pulposus with sciatica — Results of a randomized, double blind study. *JAMA* 1983; **249**: 2489-2494.

9 Fraser RD: Chymopapain for the treatment of intervertebral disc herniation — The final report of a double-blind study. *Spine* 1984; **9**: 815-818.

10 Alphen HAM van: Chemonucleolysis en lage rugklachten. *Ned Tijdschr Geneeskd* 1984; **128**: 1242-1244.

11 Deutman R: Chemonucleolysis bij patiënten met hernia nuclei pulposi lumbalis. *Ned Tijdschr Geneeskd* 1985; **127**: 1385-1390.

12 Alphen HAM van: Dokter, ik heb een hernia; kan ik een prik krijgen? *Ned Tijdschr Geneeskd* 1986; **130**: 668-670.

13 Gezondheidsraad. Commisie chemonucleolyse: *Chemonucleolyse als behandeling van de hernia nuclei pulposi.* (advies nr 1987-18). 's-Gravenhage: Gezondheidsraad, 1987.

14 Graham GE: Chemonucleolysis: a double blind study — Comparison of chemonucleolysis with intradiscal hydrocortisone in the treatment of back ache and sciatica. *Clin Orthop* 1976; **117**: 179-192.

15 Feldman J, Menkes CJ, Pallardy G, Chevrot A, Horreard P, Zenny JC, et al.: Etude en double-aveugle du traitement de la lombosciatique discale par chimionucléolyse. *Rev Rhum Mal Osteoartic* 1986; **53**: 147-152.

16 Weinstein J, Spratt KF, Lehmann T, McNeill T, Hejna W: Lumbar disc herniation. *J Bone Joint Surg (Am)* 1986; **68**: 43-54.

17 Maroon JC, Abla A. Microdiscectomy versus chemonucleolysis. *Neurosurgery* 1985; **16**: 644-649.

18 Thomas JC, Wiltse LL, Widell EH, Spencer CW, Zindrick MR, Field BT: Chemonucleolysis — A 10-year retrospective study. *Clin Orthop* 1986; **206**: 61-66.

19 Maciunas RJ, Onofrio BM: The long-term results of chymopapain chemonucleolysis for lumbar disc disease — 10-year follow-up results in 268 patients injected at the Mayo-Clinic. *J Neurosurg* 1986; **65**: 1-8.

20 Smyth H, Gallagher J, McManus F: Surgery in lumbar disc protrusion — A long term follow-up. *Ir Med J* 1983; **76**: 25-26.

21 Howe J, Frymoyer JW: The effects of questionnaire design on the determination of end results in lumbar spinal surgery. *Spine* 1985; **10**: 804-805.

22 Kardaun JWPF: *Effectiviteit en bijwerkingen van chemonucleolyse.* (achtergrondstudie, nr 1987-19) 's-Gravenhage: Gezondheidsraad, 1987.

23 Maciunas RJ, Onofrio BM: The long-term results of chymopapain — 10-year follow-up of 268 patients after chemonucleolysis. *Clin Orthop* 1986; **206**: 37-41.
24 Javid MJ: Treatment of herniated lumbar disk syndrome with chymopapain. *JAMA* 1980; **243**: 2043-2048.
25 Javid MJ: Efficacy of chymopapain chemonucleolysis. *J Neurosurg* 1985; **62**: 662-666.
26 Jabaay GA: Chemonucleolysis — 8-year to 10-year follow-up evaluation. *Clin Orthop* 1986; **206**: 24-31.
27 Watts C, Williams OB, Goldstein G: Sensitivity reactions to intradiscal injection of chymopapain during general anesthesia. *Anesthesiology* 1976; **44**: 437-439.
28 Gurdjian ES, Ostrowksi AZ, Hardy WG, Lindner DW, Thomas LM: Results of operative treatment of protruded and ruptured lumbar discs based on 1176 operative cases with 82 percent follow-up of 3 to 13 years. *J Neurosurg* 1961; **18**: 783-791.
29 Moss J, Roizen MF, Nordby EJ, Thisted R, Apfelbaum JL, Schreider BD, et al.: Decreased incidence and mortality of anaphylaxis to chymopapain. *Anesth Analg* 1985; **64**: 1197-1201.
30 Bernstein DE, Gallagher JS, Ulmer A, et al.: Prospective evaluation of chymopapain sensitivity in patients undergoing chemonucleolysis. *J Allergy Clin Immunol* 1985; **6**: 458-465.
31 Tsay YG, Jones R, Calenoff E, Sun J, Arndt L, Crispin B, et al.: A preoperative chymopapain sensitivity test for chemonucleolysis candidates. *Spine* 1984; **9**: 764-768.
32 Simmons JW, Stavinoha WB, Knodel LC: Update and review of chemonucleolysis. *Clin Orthop* 1984; **183**: 51-60.
33 Crispin BF, Stephens BG, Hansen AM, Jovero NS, Tsay YG: Clinical laboratory prediction of sensitivity reactions to chymopapain injections (chemonucleolysis). *Ann Allergy* 1986; **56**: 531.
34 Sutton, JC: Chemonucleolysis — Current status and future outlook. *Neurochirurgia* 1986; **29**: 173-178.
35 McDermott DJ, Agre K, Brim M, Demma FJ, Nelson J, Wilson RR: Chymodiactin in patients with herniated lumbar intervertebral disc(s) — An open label, multicenter study. *Spine* 1985; **10**: 242-249.
36 Spangfort EV: The lumbar disc herniation. *Acta Orth Scand* 1972; **S142**.

5

ACUTE COMPLICATIONS IN PATIENTS WITH SURGICAL TREATMENT OF LUMBAR HERNIATED DISC

Jan WPF Kardaun MD[1,2]
Lon R White MD[1]
William O Shaffer MD[3,4]

[1] Epidemiology, Demography and Biometry Program, National Institute on Aging, Bethesda, MD 20892

[2] Currently: Dept. Health Statistics, Central Bureau of Statistics, 2270 AZ Voorburg, The Netherlands.

[3] Dept of Orthopedic Surgery and Spine Surgery Section, National Naval Medical Center Bethesda, MD 20814

[4] Presently: Ventura Orthopaedic and Sports Medical Group, Ventura, CA 93003

The opinions or assertions expressed herein are those of the authors and are not to be construed as official or as necessarily reflecting the views of the Department of the Navy or of the Naval Service at large or of the Netherlands Central Bureau of Statistics.

Summary

The complications of surgical treatment for lumbar disc herniation (LDH) are important to know, but hard to measure because of their low incidence and varied pattern. Using data from the National Hospital Discharge Survey, which codes discharges and procedures according to the ICD-9-CM, we assessed acute complication rates for 3,289 surgically treated LDH patients and 4,025 nonoperative LDH patients, identifying complications from codiagnoses. The complication rates were significantly correlated with the postoperative length of stay and with the risk factors of obesity, hypertension and diabetes.

We found fewer instances of thrombophlebitis (0.3 / 1,000) and slightly lower mortality (0.9 / 1,000) than previously reported. Although the frequency of the cauda equina syndrome in the literature approximates our findings of 5 / 1,000, our data did not allow correction for the fraction of preexistent cauda equina syndromes. Our any-complication-rate is 3.7%. Even though LDH surgery is relatively safe, its complications should not be overlooked.

Keywords:

Complications / Lumbar disc herniation / Risk factors / Hospital discharge summary / Mortality.

In press, Journal of Spinal Disorders

Introduction

Lumbar herniated disc (LDH) is an endemic condition for which the effectiveness of a variety of surgical treatments has been established [1,2,3,4]. Increasing employment of percutaneous methods such as chemonucleolysis and nucleotomy has recently illuminated a need for comparative information on both effectiveness and complications, i.e. all treatment-related morbidity [5,6,7,8]. Existing data on mortality rates and complications associated with surgical treatment of LDH are sparse, having often been presented in case reports [9,10], in special studies [11,12], or as subordinate issues in studies of therapeutic effectiveness [3,13,14]. Despite the limitations of such data, the selection of treatment for an individual patient with LHD must reflect consideration of both the risk of complications and the expected therapeutic effect.

Since complications of LDH surgery are known to be both infrequent and diverse, valid estimation of rates, types, and severity of complications requires prospective data collection and very large data sets. In order to address these issues, we reviewed data from a national survey (United States) with 3,289 cases of surgically treated LDH, 4,025 cases of non-surgically treated LDH, and (for further comparison) 12,891 instances of cholecystectomy.

Patients and methods

Source of data

Data were drawn from the National Hospital Discharge Survey (NHDS), conducted annually by the National Center for Health Statistics [15,16,17]. Information from six (1980–85) annual surveys were combined for the analyses presented here. Each survey was a national probability sample of discharges from non-federal short stay hospitals in the United States. The total number of discharge records available was 1,256,293. Items considered included age, sex, dates of admission and discharge, vital status at discharge, discharge diagnoses (up to seven available), and procedures (up to four, with dates for each). For the years encompassed, diagnoses and procedures were coded according to the International Clasification of Diseases (ICD), 9th edition, Clinical Modification [18].

LDH diagnosis

Cases of LDH were identified by discharge diagnosis: ICD code 722.10 (displacement of lumbar intervertebral disc, 7,200 patients), and code 722.2 (displacement

of intervertebral site unspecified, without myelopathy, 1,015 patients) when no other diagnostic or procedural codes indicated a thoracic or cervical location.

Discectomy

A patient was considered to have undergone surgical treatment for LDH when one of the procedure codes was 805 (discectomy, 3,473 patients) and a diagnosis of LDH (see above) were present (3,289 patients with both conditions). Although this excludes patients with a negative surgical exploration for LDH, it has the advantage that patients with unrelated pathology, e.g. bone metastases, do not influence the findings. Patients with a laminectomy (ICD code 309) without discectomy (ICD code 805) were excluded from the analysis.

Nonoperative LDH patients

Patients without discectomy or laminectomy who had a first listed diagnosis of LDH (as defined above) were our primary reference group (4,025 patients). It was not thought justified to include an additional 20% patients without discectomy and an LDH diagnosis in the second through seventh diagnosis to this reference group, as the additional group includes many patients who were not primarily admitted for LDH diagnosis, evaluation, or treatment. For the purpose of the present analyses, contrast myelography, discography and spinal tap are considered non-operative treatments.

Secondary reference group

To allow comparison of the complications found after a surgical procedure that is commonly regarded as "safe" [19,20], patients with cholecystectomy (ICD code 512.2) were selected as a secondary reference group. As the only purpose of this comparison was to assure the plausibility of our findings on LDH surgery, no further refinement of this reference was performed.

Complications

Acute complications during a single admission were identified by other diagnoses listed upon discharge. A search was made for ICD codes corresponding to conditions judged likely to develop as treatment complications, particularly in association with surgical treatment of LDH. Certain of these codiagnoses could reflect either a treatment complication or a preexistent condition (as cardiac dysrythmia, aneurysm, renal failure). This problem is especially evident for conditions related to the nerve roots and spinal cord (atonic bladder, cauda equina syndrome, etc.).

In addition to being identified with conventional ICD codes, some treatment complications were separately designated with E series codes (adverse drug effects, medical misadventures, abnormal reactions) and/or with ICD codes in the 960–999 series (drug poisoning, complications of surgical and medical treatment). Codes in the range 960–966 and in the E-series were often (65%) used in conjunction with other diagnoses possibly representing a complication.

The codiagnoses were grouped into three classes (I, II, III), according to whether they were less likely, likely, and highly likely to be a complication that originated during or after surgery, based upon prima facie clinical evidence. These groupings were created *a priori* by the authors, being based on clinical judgement.

Risk factors

In addition to age, other factors that may increase the complication rate of surgical treatment of LDH were examined: obesity, diabetes mellitus, hypertension and (as a special case) tumors and/or metastases. Ascertainment of the risk factors other than age depended upon listing of the condition as a codiagnosis.

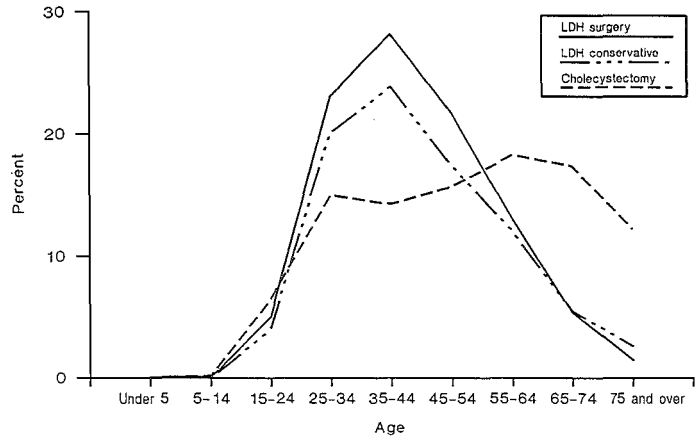
Aggregated complications

In order to summarize our findings, it was ascertained whether a patient had at least one complication in each of the classes I, II, or III. Furthermore, whether the patient had at least one complication in class II + III, or in any complication class (I + II + III). To compensate for the higher proportion of older patients in the cholecystectomy group, the age distribution of this group was adjusted to correspond with that of the LDH surgery group (indirect standardization).

Results

The numbers of patients per year in the study group and the reference groups are presented in Table 1. It appeared that a moderate but consistent increase in the number of admissions for LDH (both for patients with surgery and nonoperative treatment) occurred between 1980 and 1985, whereas the admissions for cholecystectomy rose far less. The total number of actual admissions of patients with surgical treatment for LDH was 3,289, with a nonoperative reference group of comparable size and a cholecystectomy reference group of fourfold size (12,891). The age distributions of the LDH patients with surgical treatment and nonoperative treatment were quite comparable (Fig. 1). The cholecystectomy group did not have a peak at 35–44 years, and had more patients of 65 and over, a fact that must kept in mind when the complication rates are compared.

Figure 1 — Age distribution of patients with surgical or non-surgical treatment of lumbar disk herniation or with cholecystectomy, National Hospital Discharge Survey, 1980 - 1985.



The codiagnoses and complications noted on the discharge summary are presented in Table 2 for patients with discectomy (and a diagnosis of LDH) as the primary study group, for LDH patients without surgery and for patients with a cholecystectomy.

Table 1 — Number of patients in the NHDS sample with surgical or non-operative treatment for lumbar disk herniation or cholecystectomy.

Year	total NHDS sample	LDH, surgical N	LDH, surgical per 1000 admissions*	LDH, non-surgical N	LDH, non-surgical per 1000 admissions	cholecstectomy N	cholecstectomy per 1000 admissions
1980	223,785	454	2.0	575	2.6	2,192	9.8
1981	226,585	452	2.0	622	2.8	2,232	9.9
1982	213,732	506	2.4	644	3.0	2,196	10.3
1983	206,027	544	2.6	673	3.3	2,035	9.9
1984	192,083	607	3.2	740	3.9	2,038	10.6
1985	194,081	726	3.7	771	4.0	2,198	11.3
Total	1,256,293	3,289	2.6	4,025	3.2	12,891	10.3

* I.e. the number of admissions per 1000 in the same diagnostic category and year. Applying the sampling weights to each patient did not result in noteworthy differences for the rates in this table.

Table 2 — Complications of surgical and non-operative treatment for LDH and of cholecystectomy, National Hospital Discharge Data, 1980-1985.

complications, based on ICD-9-CM	LDH surgery		LDH non-operative		Cholecyst-ectomy	
	N	per 1000 admiss.	N	per 1000 admiss.	N	per 1000 admiss.
class I — less likely to be a complication of treatment						
conduction disorders	4	1.2	7	1.7	69	5.4
cardiac dysrythmias	12	3.6	22	5.5	338	26.2
heart failure	3	0.9	10	2.5	181	14.0
aortic aneurism	0	0.0	2	0.5	67	5.2
other aneurysm	0	0.0	0	0.0	11	0.9
acute CVA; chronic ischemia cerebri	0	0.0	0	0.0	2	0.2
renal failure unspecified	1	0.3	0	0.0	29	2.2
atonic bladder	0	0.0	2	0.5	6	0.5
other functional disorder bladder	2	0.6	0	0.0	0	0.0
retention of urine	6	1.8	9	2.2	49	3.8
incontinence of urine	3	0.9	1	0.2	8	0.6
class II — likely to be a complication of treatment						
septicemia	0	0.0	1	0.2	105	8.1
meningitis, encephalitis	12	3.6	11	2.7	1	0.1
acute myocardial infarction	3	0.9	1	0.2	64	5.0
(thrombo)phlebitis	1	0.3	7	1.7	28	2.2
other venous embolism	1	0.3	1	0.2	9	0.7
subarachnoid hemorrhage	0	0.0	0	0.0	2	0.2
transient ischemic attack	3	0.9	2	0.5	18	1.4
acute & ill defined acute CVA	0	0.0	2	0.5	30	2.3
acute edema of lung	0	0.0	0	0.0	4	0.3
acute renal failure	0	0.0	0	0.0	24	1.9
class III — highly likely to be a complication of treatment						
acute pulmunary embolus	1	0.3	3	0.7	47	3.6
arterial embolism	0	0.0	1	0.2	23	1.8
postlaminectomy syndrome, cauda equina s.	26*	7.9	12	3.0	1	0.1
drug poisoning	0	0.0	0	0.0	2	0.2
certain adverse reactions	5	1.5	13	3.2	37	2.9
complications of med. & surg. care, NEC	42	12.8	10	2.5	419	32.5
misadventures	0	0.0	0	0.0	1	0.1
abnormal reactions	43[¶]	14.9	8	2.0	391	30.3
adverse drug reactions	11	3.3	22	5.5	70	5.4
death	3[†]	0.9	2	0.5	187	14.5

* postlaminectomy 12, cauda equina syndrome 16, both listed 4. See text for special status of cauda equina syndrome. [†] 1 myocardial infarction; 1 mycocardial infarction with cardiac arrest; 1 peripheral vascular anomaly. One aged 55-64, two aged 65-74. [¶] among them 4 central nervous system-, 4 respiratory -, 3 G.I. complications, 3 hemorrhages, 4 accidental punctures, 1 foreign body, 1 persistent fistula, 10 other specified complications NEC.

The class I complications occurred as often or more often in the reference groups than in the LDH surgery group except for incontinence of urine or other functional disorders of the bladder.

Among class II complications transient ischemic attacks (TIAs) and meningitis/encephalitis occurred more often in LDH surgery patients than in nonoperative LDH patients. Meningitis was almost absent in the cholecystectomy group, while TIAs occurred in this group more often than in both LDH groups.

In the class III complications, there was a stronger difference between the surgical and nonoperative LDH group, based upon larger numbers. The codes for drug poisoning and medical and surgical misadventures hardly occurred. Adverse drug reactions and other adverse reactions were less frequent in the LDH surgery group than in the reference groups, but together they occurred in almost 1 per 200 admissions. Abnormal reactions and complications of medical and surgical care both occurred much more frequently in LDH surgery patients than in nonoperative LDH patients, though about half as frequent as in the cholecystectomy group.

Table 3 Aggregate numbers of complications.

Complication class	LDH, surgery		LDH, no surgery			Cholecystectomy crude		Cholecystectomy age adj.		
	N	per 1000 admiss.	N	per 1000 admiss.	p_1	N	per 1000 admiss.*	N	per 1000 admiss.	p_2
I	31	9	50	12	0.13	680	53	286	22	<0.001
II	20	6	25	6	0.53	269	21	110	9	0.10
III	82	25	52	13	<0.001	721	56	491	38	<0.001
II + III	97	30	73	18	<0.001	889	69	561	43	<0.001
I + II + III	123	37	121	30	0.05	1,343	104	766	59	<0.001

age adj. = the age distribution of cholecystectomy patients is adjusted to that of LDH surgery patients.

p_1 and p_2 are the probabilities that the difference in complications (on a row) between LDH surgery patients and non-operative LDH patients and age adjusted cholecysctectomy patients, respectively, are statistically significant.

* the difference of crude complications rates for cholecystectomy was in all cases significantly higher than for LDH surgery ($p < 0.001$).

For the meaning of class I, II, III see Table 2. The numbers in the seperate classes do not add up to the numbers in the combined classes, because a patient might have had both a class I and a class II complication for example.

The diagnoses of postlaminectomy syndrome or a cauda equina syndrome occurred almost exclusively in LDH patients, more in surgical than nonoperative patients. The character of a cauda equina syndrome is ambiguous, as it can be both a consequence of LDH surgery and an indication for it. The data source used for this analysis does not allow the resolution of this ambiguity.

A special case among the complications is occupied by mortality; almost 1 death occurred per 1000 admissions for surgical LDH patients, a rate double that of nonoperative LDH patients but only a fraction of the mortality associated with cholecystectomy.

In considering the co-occurrence of single diagnoses, as in Table 2, some patients were counted more than once if more than one complication was listed, e.g. for dysrythmias and conduction disorders or for acute myocardial infarction and death. Codes for "complication of medical and surgical care" and "abnormal reactions" were used often in conjunction with other codes.

The aggregated complications are shown in Table 3. In all three patient groups, the complication rate was lowest in class II, and highest in class III. The class I complication rate was especially high for nonoperative LDH patients and cholecystectomy patients (not age adjusted). Combining classes II + III resulted in almost the same relative complication rates (between the patient groups) as class III only. The complication rate in class II + III was statistically higher for the LDH surgery group than for the nonoperative LDH group, but significantly lower than for the cholecystectomy group after age adjustement.

Obesity, hypertension, and diabetes are commonly viewed as risk factors for the development of surgical complications. Their prevalence in the three patient groups is shown in Table 4. There were fewer patients with these risk factors in the LDH surgery group than in the nonoperative LDH group, perhaps reflecting a slight reluctance to operate upon these patients. The cholecystectomy group had

Table 4 — Selected risk factors for a complicated course of surgical treatment for LDH.

Risk factor	LDH, surgical		LDH, non-surgical		Cholecystectomy	
	per 1000		per 1000		per 1000	
admissions						
Obesity	37	11.2	102	25.3	531	41.2
Hypertension	100	30.4	192	47.7	887	68.8
Diabetes	63	19.2	107	26.6	614	47.6
Tumors (all)	7	2.1	15	3.7	489	37.9
with metastases	1	0.3	8	2.0	139	10.8

Table 5 — Length of stay and risk factors related to the number of class II + III complications in patients with surgical treatment for LDH.

Factor	Spearman's r	Mantel-Haenszel Odds Ratio	p
Length of stay	0.11		$\ll 0.001$
pre-operative	0.02		0.12
post-operative	0.10		$\ll 0.001$
Age	0.03		0.07
Obesity		4.1	0.02
Hypertension		3.5	0.002
Diabetes		2.9	0.04
Tumors, incl metas.		0.0	1.0

the highest rates for all these conditions, probably related to the difference in age distribution (see below).

Among surgically treated LDH cases, the importance of all three risk factors was supported by statistically significant associations with development of complications (defined as above) during the course of hospitalization (Table 5). Codiagnoses of obesity, hypertension or diabetes were three to four times as common among patients having a complicated course. In contrast, no significant association was found between complications and neoplastic disease, including metastatic tumors. Advancing age was only marginally significantly associated with the development of complications of treatment ($p = 0.7$, Spearman's rank correlation).

As expected, the length of hospitalization for LDH surgery admissions with complications (likely or highly likely) was significantly longer than that for other surgically treated LDH cases. The association with complications was statistically significant ($p < 0.001$) only for length of the post-operative interval, as opposed to the pre-operative interval ($p = 0.12$). This substantiates our assumption that most of these conditions developed or worsened during or after the surgical procedure.

Data from the literature

Only a few reports have been published with emphasis on complications of LDH surgery [3,21,22,23]. Many other reports mention only a few types of complications,

either by special focus [24,25] or because of limited numbers. Still, there are many papers that can be used as data source for complication rates of LDH surgery, summarized in Table 6.

One large multicenter study in the Federal Republic of Germany has reported very high complication rates, such as 115 / 1,000 spondylitis and 11 / 1,000 instances of acute pulmonary embolism. Because these numbers are so divergent from the remaining literature, we have excluded them from the ranges given below.

Most authors mention wound infections (6–47 / 1,000) and mortality (0.2–3 / 1,000). Some authors mention thrombophlebitis (2–17 / 1,000). A variety of other infectious complications are described, such as meningitis, discitis, spondylitis, osteomyelitis and — less specific — urinary tract infections. "Misadventures" include neurological damage, root dissections, and perforation of the ventral major vessels. The occurrence of shock (4 / 1,000) is of special interest, because of the important role of shock in the discussion about the safety of chymopapain [8]. The cauda equina syndrome is reported to occur as a complication in 2–3 patients / 1,000.

Some less common or less completely documented complications, not mentioned in the table, are dislocation of vertebra, meningocele, fistula, dura mater tears mentioned without numbers [23]; (a few) dura mater tears [13]; major vessel injuries (106 complications, 3,000 physicians questioned) [12]; two ulnar nerve neuritis (out of 905 patients) [14]; one hemiplegia (out of 159 patients) [26]; one paralysis of quadriceps [26]. An old but clear description of vascular injuries does not give precise numbers [12].

Discussion

An important factor influencing the choice of a specific treatment for lumbar herniated disc is a detailed knowledge of what complications are to be expected from that treatment. The use of a national representative database should result in a realistic estimate of complications for average quality of care, including errors in act and judgement that inevitably occur. Such an estimate would not apply to ideal circumstances, ideal physicians and ideal patients. In this paper, the term "complication" is used in a broad sense, i.e. any unintentional* negative effect on the patient that is not due to ineffectiveness of therapy and that would not have occurred to the patient if the treatment would not have taken place. The

---

* This is to allow certain undesirable but unavoidabe side effects of a treatment not to be called complications. E.g. the functional impairment after a foot amputation for diabetic gangrene is not a complication.

Table 6 —Complications of LDH surgery mentioned in the literature

complication	1st author [ref]	year	size of study*	per thousand
Infectious				
wound infections	Spangfort[3]	1972	2,504	32[1]
	Spangfort[3]	1972		6[2]
	(Spangfort[3]	1972)[†]	10,104	29[3]
	Gurdjian[13]	1961	1,176	112
	Schepelman[21]	1977	1,645	6 to 200
	Stevens[25]	1964	154[6]	39
	Aitken[26]	1952	158	30
	Nachlas[27]	1952	374	47
	Barr[28]	1967	220[6]	9
	Horwitz[24]	1975	496	30[4]
	Raaf[14]	1959	905	8
	Wright[11]	1970	670	31
meningitis	Schepelman[21]	1977		2
discitis	Spangfort[3]	1972		20
spondylitis	Spangfort[3]	1972		0.6
osteomyelitis	Oppel[22]	1977	3,032	51
	Gurdjian[13]	1961		5[5]
urinary tract infections	Schepelman[21]	1977		94
pneumonia	Schepelman[21]	1977		6
Circulatory				
shock	Nachlas[27]	1952		4
	Raaf[14]	1957	905	6
acute pulmonary embolus	Schepelman[21]	1977		11
	Raaf[14]	1952		2
	Aitken[26]	1952		19
thrombophlebitis	Schepelman[21]	1977		15
	Spangfort[3]	1972		2
	(Spangfort[3]	1972)	6,385	17[3]
	White[29]	1966	159	18
	Barr[28]	1967		14

Table 6 (Continued)

Surgical mishap				
neurological damage	Barr[28]	1967		14
	Munro[30]	1956	375	11
	White[29]	1966		20
roots dissected	Munro[30]	1956		8
ventral perforations	Oppel[22]	1977	3,038	6
cauda equina syndrome	Oppel[22]	1977		3[7]
	Spangfort[3]	1972		2[8]
	Aitken[26]	1952		19
bladder paralysis	Munro[30]	1956		45
Death				
	Mayfield[23]	1976	1,408	3[9]
	Oppel[22]	1977		3
	Spangfort[3]	1972		1
	(Spangfort[3]	1972)	22,888	3[3]
	Munro[30]	1956		11
	White[29]	1966		3
	Roberts[6]	1986	15,000	0.2
	Schultz[32]	1958	4,000	0

* The study size is given only for the initial entry of a publication in this table.
† The study by Spangfort contains also data that stem from other sources; indicated with brackets.

NOTES: [1] moderate wound infections, [2] severe wound infections, [3] compilation of other literature,[4] perioperative antibiotics, [5] six weeks postoperative, [6] excluding patients with fusion,[7] in repeat operations, [8] see also text of [30], [9] some tumors and metastases

term "complication" does not carry a connotation of fault, error or negligence or suboptimal care throughout our article, though these may increase the number of complications. The lack of effectivity is regarded by other investigators sometimes as a complication ("failed back") [22], but in our opinion the measurement of effectivity should be considered a separate issue from that of complications.

The National Hospital Discharge Survey dataset has several advantages for assessing complication rates. The data are collected in a prospective and unbiased fashion, on a sample that is representative for the USA, providing larger numbers than most clinical series. It is easy to define reference groups and to examine codiagnoses for risk factors. First however, the primary disadvantage is that complications cannot always be distinguished from preexistent pathology. Second,

biased or selective recoding of codiagnoses cannot be ruled out. Although dates of all surgical procedures are given, this dataset provides neither a date of onset for diagnoses nor means for distinguishing first admissions from repeat admissions of the same patient. These problems limit the use of this data set to acute complications developing during a single admission. Late complications, such as scarring, spinal instablility, secondary stenosis, may escape our attention or may even be counted towards the nonoperative LDH group on occasion of a later admission. The nonoperative LDH group cannot be sharply defined, because strict bedrest is not a recognizable ICD procedure. Despite these limitations, the NHDS dataset offers a convenient way to approach questions on complication rates.

Randomized clinical trials are of limited use in assessing complication rates. The large numbers required would make such a study impractical and unethical. Furthermore, randomized controlled trials for assessing effectivity often impose limits on the age of the patients [33,34,35], require unusually strict entry criteria, and can do with a few hundred patients. All these factors limit their usefulness for assessing complication rates. We compared the complications of LDH surgery patients with those of nonoperatively treated LDH patients and other surgical non-LDH patients (cholecystectomy), realizing that there may be many risk factors known to the surgeon-in-charge, that may have influenced the decision to operate or not. Part of this selection bias was overcome by considering only nonoperative patients with a *main* diagnosis of LDH.

The overall complication rate in our study (30 patients / 1,000) is rather high. While this may be a consequence of our method of assessment, it may well be realistic. Nonoperative treatment for LDH (18 / 1,000) and cholecystectomy, even age adjusted to the LDH population (43 / 1,000), had high complication rates. Apparently a hospital is not a healthy place to stay.

The mortality was 0.9 / 1,000, which is higher than has been reported for chemonucleolysis (0.6 / 1,000 [36], 0.4 / 1,000 [31], 0.2 / 1,000 [6]). The difference between surgical and nonoperative treatment is significant. A comparison of these groups is limited by the possible selective attribution of high risk patients to the nonoperative group.

In more detail, the categories *likely* and *highly likely* complications (class II and II, resp.) appeared most useful; the *less likely* complications (class I) occurred more often in nonoperative patients. We found less thrombophlebitis (0.3 / 1,000) than in the literature (2–15 / 1,000). The cauda equina syndrome (4.9 / 1,000), for which we could not assess a prior or post-operative status, was comparable to the literature (3–19 / 1,000). The cauda equina syndrome caused by a disc herniation is well documented [37,38]. The only published report to consider pre-operative conditions with relation to post-operative conditions attributed 17 cases of bladder paralysis mostly to postoperative cauda compression, whereas only one case had a similar condition preoperatively [30]. Some other complications, notably wound

infections, discitis and spondylitis are difficult to recognize using ICD codes, as is the case with root dissections, neurological damage and ventral perforations of the disc.

In the literature quite a variation in complication rates and types of complications exists, partly due to small samples and lack of a common definition of complications. Sometimes, rates in the literature may contain also late complications. For individual complications, the incidence is so low, that due to the small numbers statistical significant differences cannot be established.

In this large and national representative dataset, we found a varied list of complications for surgical treatment of Lumbar Herniated Disc, that amounted to a higher number of complications than in most other reports. Reports on only a few hundred patients do not give good insight in the complications that can be expected. These facts have to be considered when alternative treatments, such as chemonucleolysis and percutaneous nucleotomy are evaluated.

References

1 Weinstein J, Spraat KF, Lehmann T, McNeil T, Hejna,W: Lumbar disc herniation. A comparison of the results of chemonucleolysis and open discectomy after ten years. *J Bone Joint Surg [Am]* 1986; **68A**: 43-54.

2 Hudgins WR: Computer aided diagnosis of lumbar disc herniation. *Spine* 1983; **8**: 604-15.

3 Spangfort EV: The lumbar disc herniation. *Acta Orthop Scand*, **S142**, 1972.

4 Hakelius A: Prognosis in Sciatica. *Acta Orthop Scand* 1970, **S129**.

5 LeBlanc FE: Sciatica Management by chenonucleolysis versus surgical discectomy. *Neurosurg Rev* 1986; **9**: 103-107.

6 Roberts MP: Mortality Rate of lumbar discectomy for herniated intervertebral disc disease. Presented at the *Annual Meeting of the American Association of Neurologic Surgeons*, San Francisco, CA, 1984, p 69.

7 Sutton JC: Chemonucleolysis - current status and future outlook. *Neurochirurgia* 1986; **29**: 173-178.

8 Kardaun JW, Schipper J: [Chemonucleolysis for lumbar nucleus pulposus hernia; efficacy and side effects.] *Ned Tijdschr Geneeskd* 1988; **132**: 285-289

9 Harbison SP: Major vascular complications of intervertebral disc surgery. *Ann Surg* 1954; **140**: 342-348.

10 Suttorp MJ, Weigel HM, van der Wolk DB, Veraart BE: Severe congestive heart failure and clotting abnormalities after lumbar disc surgery. *Neth J Med* 1988; **32**: 84-89.

11 Wright RL: *Septic complications of neurosurgical spinal procedures.* Springfield, IL, Charles C Thomas, 1970.

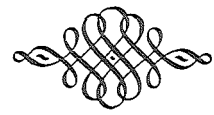
12 DeSaussure RL: Vascular injury coincident to disc surgery. J *Neurosurg*, 1959; **16**: 222-228.

13 Gurdjian ES, Ostrowski AZ, Hardy WG, Lindner DW, Thomas LM: Results of operative treatment of protruded and ruptured lumbar discs. *J Neurosurg*, 1961; **18**: 783-791.
14 Raaf J: Some observations regarding 905 patients operated upon for protruded lumbar intervertebral disk. *Am J Surg* 1959; **97**: 388-397.
15 National Center for Health Statistics: *Public Use Data Tape Documentation; National Hospital Discharge Survey, 1979.* National Center for Health Statistics, Hyattsville, MD, 1981.
16 National Center for Health Statistics, WR Simmons: *Development and the design of the NCHS Hospital Discharge Survey.* Vital and Health Statistics, series 2, no 39, PHS pub no 1000. Public Health Service, Washington, DC, US Government Printing Office, 1970.
17 Graves EJ: *National Hospital Discharge Survey: Annual Summary 1987.* Vital and Health Statistics, series 13, no 99. DHHS pub no (PHS) 89-1760. National Center for Health Statistics, Hyattsville, MD, 1989.
18 National Center for Health Statistics: *International Classification of Diseases, 9th Revision, Clinical Modification.* DHHS pub no (PHS) 80-1260. Public Health Service, Washington, US Government Printing Office, 1980.
19 McSherry CK, Glenn F: The incidence and causes of death following surgery of nonmalignant biliary tract disease. *Ann Surg* 1980; **191**: 271-275.
20 Shamma'a J: Shockwave lithotripsy of gallbaldder stones [letter to the editor regarding Sackmann M et al., New Engl J Med, 1988: 318: 393-397.] *New Engl J Med* 1988; **319**: 371-372.
21 Schepelmann F, Greiner L, Pia HW: Complications following Operation of herniated lumbar discs. *Adv Neurosurg* 1977; **4**: 53-54.
22 Oppel F, Schramm J, Schirmer M, Zeitner M: Results and complicated course after surgery for lumbar disc herniation. *Adv Neurosurg*, 1977; **4**: 36-51.
23 Mayfield FH: Complications of laminectomy. *Clinical Neurosurgery*, 1976; **23**: 435-439.
24 Horwitz NH, Curtin JA: Prophylactic antibiotics and wound infections following laminectomy for lumbar disc herniation. A retrospective study. *J Neurosurg*, 1975; **43**: 727-731.
25 Stevens DB: Postoperative orthopedic infections. A study of etiological mechanisms. *J Bone Joint Surg [Am]* 1964; **46**: 96-102.
26 Aitken AP: Rupture of the intervertebral disc in industry. *Amer J Surg* 1952; **84**: 261-267.
27 Nachlas IW: End-result study of the treatment of herniated nucleus pulposus by excision with fusion and without fusion. *J Bone Joint Surg [Am]* 1952; **34A**: 981-988.
28 Barr JS, Kubik CS, Molly MK, McNeill JM, Riseborough EJ, White JC: Evaluation of end results in treatment of ruptured lumbar intervertebral discs with protusion of nucleus pulposus. *Surg Gynecol Obstet* 1967; **125**: 250-256
29 White JC: Results in surgical treatment of herniated lumbar intervertebral discs. *Clin Neurosurg* 1966; **13**: 42-54.

30 Munro D: Lumbar and sacral compression radiculitis. *New Engl J Med* 1956; **254**: 243-252.
31 Agre K, Wilson RR, Brim M, McDermott DJ: Chymodiactin post-marketing surveillance - demographic and adverse experience data in 29,075 patients. *Spine* 1984; **9**: 479-485.
32 Schultz EC: Postoperative bone changes following lumbar disc removal. *J Neurosurg* 1958; **15**: 537-547.
33 Fraser RD: Chymopapain for the treatment of intervertebral disc herniation -the final report of a double-blind study. *Spine* 1984; **9**: 815-818
34 Javid MJ, Nordby EJ, Ford LT, Hejna WJ, Whisler WW, Burton C, Millet K, Wiltse LL, Widell EH, Boyd RJ, Newton SE, Thisted R: Safety and efficay of chymopapain (Chymodiactin) in herniated Nucleus pulposus with sciatica. *JAMA* 1983; **249**: 2489-2494.
35 Van Alphen HA, Braakman R, Bezemer PD, Broere G, Berfelo MW, Kostense PJ: [Chemonucleolysis or herniotomy; results from a comparative study in patients with lumbar nucleus pulposus hernia] *Ned Tijdschr v Geneeskd* 1988; **132**: 304-308
36 Watts C: Complications of chemonucleolysis for lumbar disc disease. *Neurosurgery* 1977; **1**: 2-5.
37 Kostuik JP, Harrington I, Alexander D, Rand W, Evans D: Cauda equina syndrome and lumbar disc herniation. *J Bone Joint Surg [Am]* 1986; **68A**: 386-91.
38 Scott PJ: Bladder paralysis in cauda equina lesions from disc prolapse. *J Bone Joint Surg [Br]* 1965; **47B**: 224-35.

6

TOWARDS A RATIONAL DIAGNOSIS AND TREATMENT OF LUMBAR DISK HERNIATION



I Introduction

"One of the most formidable problems which has long faced the medical profession has been the treatment of low back pain. ... Cognizance of [the] inefficiency [of treatment] has led the medical profession to accept with enthusiasm any new development which has promised to be of assistance." In the more than 40 years that have elapsed since these statements, cited from Aitken [2], have been published, their relevance has hardly deteriorated.

Nowadays, however, there is a tendency to evaluate critically the assets that new procedures in diagnosis or treatment can bring us, based on a methodology that matured in the past decade to analyze problems of clinical protocols [39]. By combining data on the quality and the side-effects of diagnostic procedures, and on the effectiveness and complications of treatments, the expected advantage of one course of diagnostic and therapeutic decisions can be calculated. Such analyses provide guidance, but nothing more, in making these decisions. If an analysis shows clearly that there are not many advantages nor disadvantages in using a certain test or treatment, then a vehement debate appears hardly worthwhile.

The present study aims at performing an analysis of two diagnostic modalities (CT and myelography) and three therapeutic modalities (non-surgical treatment, (wide) laminectomy plus discectomy, and chemonucleolysis) in patients that have a clinically suspected lumbar disk herniation (LDH).

II Overview of the problem setting

Diagnosis and treatment of lumbar herniated disk has many aspects; one could focus on anatomy for instance, or on operative technique or on rehabilitation methods. In order to be accessible for semi-quantitative research, a reduction in the complexity of the clinical situation of diagnosis and treatment of LDH is required to formulate our research problem.

a. History

The analysis concerns patients, aged 18–65, with symptoms and complaints that point to an LDH. Typically, these might be pain, radiating from the low back into the buttock and/or leg, positive Straight Leg Raising test (Lasègue), exacerbation of the pain on coughing and sneezing, etc. The patient may have had treatment with strict bedrest — to no avail. Prior low back surgery is supposed not to have been undergone, and no radiographs, except for plain x-rays, have been taken. The patient is referred to the outpatient department of a hospital, e.g. to a neurosurgeon or an orthopedist — henceforth generally denoted by "surgeon" —, with the

question to confirm or reject the diagnosis of LDH and to consider the patient for operation.

b. Investigations

The surgeon will of course obtain a new history from the patient and perform a physical examination. Except for observer variation and quality differences in judgement, this should not reveal new information, as it is a repeated measurement. [*Note 1*] The investigations considered are myelography and spinal CT [5, 8, 11, 23, 47]. Other radiological investigations like discography and phlebography are not in focus nowadays. Phlebography, even though it has become outdated, will be used to illustrate important points, which are not (yet?) visible by studying the history of CT. Sufficient data are not yet available to include MRI in the analysis. Nonetheless, it is possible to calculate the minimum quality that MRI should prove to have in future, in order to surpass the qualities of the established procedures. Ideally, a complication rate should be known for each procedure considered.

c. Treatments

Given the diagnosis that results from clinical and radiological examination, the following options are essentially available: Non-surgical treatment, chemonucleolysis, nucleotomy. Variations like percutaneous nucleotomy [16] and micro-nucleotomy [44] are not considered in this chapter. Modifications of non-surgical treatment, e.g. aggressive physical rehabilitation [52] will not be considered separately, as will surgery after a fruitless CNL [10, 29]. Each of the procedures involves risks of complications, which are condensed into a few categories.

d. Outcomes

Regarding the effectiveness of treatment, four states are usually distinguished: completely improved, improved but not pain-free, no change, or worse than before. Sometimes three states (excellent, good, poor) or even two states (success, failure) are used. Such simplification fails in describing aspects that are each normally relevant in a small fraction of the patients, such as recovery of motor loss [6], or the return to the previous job. It may be reasonably assumed that these patients are proportionally included in the many studies that report on outcome in categories similar to the above [34].

Complications, which occur only in a small fraction of the patients may be categorized into five classes of decreasing severity like death, disability, permanent impairment, temporary impairment, protracted stay in hospital.

e.1. Utilities — physician

Ideally, the personal and professional values of the physicians should yield for those of the patient. The physician is supposed to offer his expert knowledge and to allow the patient to give his (i.e. the patient's) evaluation of the possible benefits (or lack of these) of the proposed treatment [65]. In practice, however, this is not completely feasible. One of the instances where the physician's expert judgement and his implicit values are diffficult to separate is in his diagnostic evaluation of radiological images. As can be shown empirically [36, 37], physicians tend to differ in their preference for a false negative or false positive diagnosis.

e.2. Utilities — patient

There are as yet no empirical data to assess the values which (most) patients would attribute to the outcome states. The categorization of the variations in outcome and complications into a few results allows, in principle, to weigh these with values that seem reasonable to those who are not personally involved in the choice. For individual patients quite different utilities are thinkable, which influence the preference of the choices accordingly.

III Investigations

a. Assumptions

Considerable attention has been given in the literature to the assessment of the relative qualities of radiological diagnostic procedures for LDH. Not all published papers appear to be useful in retrospect [36]. If three assumptions are made, viz

(1) the "real" presence of LDH can be accurately assessed during spinal surgery,
(2) the patient selection for surgery is not influenced by the radiologic findings,
(3) the clinical findings do not influence the radiological diagnosis,

some findings can be distilled from the literature. [*Note 2*]

The findings of CT and myelography are *not* assumed to be independent, even if the two images are judged by independent observers, as it is obvious that a patient with a positive CT will be more likely than a patient with a negative CT to have a positive myelography.

b.1. Measures — odds ratio (OR)

The measure of discriminatory power of radiological procedures used in this chapter will be the odds ratio (of a radiological and surgical diagnosis of LDH). One description of this odds ratio may be: the odds of radiologically detecting an LDH in patients with true LDH divided by the odds in patients without a true LDH. [*Note 3*]

Based on a number of publications, estimates can be made for the odds ratios of myelography (OR= 19 (95% C.I.= 14-27)), and CT (OR=12 (95% C.I.= 7-21)) [37]. Assuming a (realistic) estimate of 80% prior probability of LDH [36] in all patients submitted to myelography or CT throughout this report, and that false negative errors (FNE) and false positive errors (FPE) [*Note 4*] are to be avoided

Table 1 – Test qualities for several Odds Ratios and Threshold Values for a prior probability of 0.80

	OR	Log(TV)	sens	spec	c.e.r.
a)					
	12	0	0.78	0.78	0.22
		2	0.90	0.56	0.16
		3	0.94	0.44	0.16
		5	0.98	0.22	0.17
b)					
	19	0	0.81	0.81	0.19
		2	0.92	0.62	0.14
		3	0.95	0.49	0.14
		5	0.98	0.26	0.16
c)					
	60	0	0.89	0.89	0.11
		2	0.95	0.74	0.09
		3	0.97	0.63	0.10
		5	0.99	0.39	0.13

The test qualities sensitivity (sens), specificity (spec), and crude error rate (c.e.r.) are calculated for three values of the Odds Ratio (OR) and four values of the Threshold Value (TV).

to an equal extent, this would result in a sensitivity (equal in magnitude to the specificity) of 81% and 78%, respectively. (See Table 1a) Also the difference in the crude error rate is rather small (19% and 22%, resp.).

b.2. Measures — threshold value (TV)

The threshold value (TV) is the usual term for the border point of a certain test between diseased and normal cases. A neutral threshold value corresponds to a situation with as many FPE's as FNE's. In ROC space, studies with a neutral TV lie on the 135 diagonal.

Empirically, it appears that false positive errors are not always valued equally "bad" as false negative errors (see fig. 2 in [37] (*p. 56 in this book*)). There is considerable variation. Some papers give results corresponding to a situation where a FPE is considered worse (at least, occurs less frequently) than a FNE. This can be seen from the fact that the corresponding confidence ellipses are to the left of the 135 diagonal. However most show the opposite situation (again, see Fig. 2 of [37] (*p. 56*)). Expressed as the log(TV), i.e. the distance of the center of the ellipse to this diagnonal [*Note 5*], both for CT and myelography, the average preference of FPE over FNE is 2–3, while 0 and 5 are realistic extremes. [*Note 6*]

An increase in test quality has a different effect at a neutral TV (i.e. Log(TV) = 0) than at a clearly non-neutral TV. Using, for example, the same estimates for the OR for CT and myelography as before (in *b.1*), but now applying a log(TV) of 2 and 3 (see Table 1b and 1c), it appears that the sensitivity has improved at the expense of a deteriorated specificity, while the error rate improves. In other words, considering only the error rate improvement does not account for the fact that the test qualities remained constant, nor that this error rate improvement is reached at the expense of a low specificity.

b.3. Conclusions

The above can also be interpreted as demonstrating that an increase of test quality (OR) is less effective at a log(TV) of 2 - 3 than at a log(TV) of 0. Suppose a new or future test, such as MRI, would have a "real" OR of 60, then we can see from Table 1b and c, that it will be difficult, though not impossible [*Note 7*] to assess this difference between an OR=60 and an OR=19. At this moment, we must recognize the constraints, and admit our limited ability to assess the difference in sensitivity of 92% and 95% and in specificity between 62% and 74% (O.R. = 19 and 60, resp., for log(TV)=2, Table 1) if we use surgical inspection as a criterion for the presence of an LDH and if only part of the test-negative patients are sent in for surgery.

In other words, in the TV range that is met most often in daily practice of LDH radiology, a gain in test quality is less useful than at a neutral TV. Moreover, our

golden standard (surgical verification of LDH) is limited in quality and is not available for all patients. Thus, paradoxically, even if a superior technique were to become available, we will perhaps never know, for lack of a suitable measuring instrument.

Another conclusion emerging from section b.2, is that a difference in TV, which is more a property of the observer than of the radiological procedure, can be as important as a difference in OR. Therefore, researchers reporting on radiological procedures, as well as practicing clinicians, need to be aware of their position on the TV axis. In practical situations equal values of FPE and FNE cannot always be defended. For research papers that focus on the test qualities, it is advisable to require that observers aim at an equal sensitivity and specificity, or at least describe why they did otherwise.

c.1. Two tests — interaction

Until now, only one radiological test at a time was considered. It is of course posssible to combine two (and three) radiological tests in the expectation that this combination gives better results. If myelography and CT would give truly independent assessments of the presence of an LDH, a combined CT and myelography would give better results, if the posterior (or *post test*) probabilities would be the criterion. [*Note 8*] The combined test result has to be dichotomized, however, as the decision to operate is a *yes or no* decision, hereby sacrificing some information. In Table 2 some data on combined CT and myelography are summarized from Schipper et al. [55] [*Note 9*]

It is conceivable that CT can visualize exactly those LDH's that myelography would miss, because of the difference between the section or projection that is viewed, and the other differences in imaging. Such issues have been considered in the literature on a case by case basis [17, 53]. On the other hand it is conceivable that "difficult backs" are hard to diagnose by either technique. It is even possible that CT and myelography would make exactly the same mistakes, and would merely support each other's correct and incorrect diagnoses.

These hypotheses can be tested by a 2×2×2 table, as in Table 2a. [*Note 10*] If myelography (MY) and CT were independent, the overall Odds Ratio for CT would be the same as for the two strata with MY+ and MY–, and equivalently the overall Odds Ratio for MY would be the same as for the two strata CT+ and CT–. This is clearly not the case (in statistical terms, there is a strong interaction between CT and myelography results).

It appears that one negative test implies that the other test will have less diagnostic efficacy as expressed by the odds ratio. In other words, it is more difficult to establish the absence of an LDH, than the presence. This is in accordance with

Table 2 – Combination of Myelography and CT, derived from Schipper et al, 1987 [55].

a)

The 2 × 2 × 2 table gives the frequencies in 195 patients for all combinations of myelographic (MY), computed tomographic (CT) and surgical (LDH) diagnosis.

	MY+		MY–	
	CT+	CT–	CT+	CT–
LDH+	101	35	19	10
LDH–	1	9	6	14

b)

Stratified Odds Ratios can be derived from a) by considering the subgroups CT+ and CT– separately for the myelography OR and vice versa. The stratified OR's indicate that there exists a strong interaction between myelography and CT.

MY	all	9.4	CT	all	8.8
	CT+	31.9		MY+	26.0
	CT–	5.4		MY–	4.4

c)

Combined test results can be derived from a) by dichotomizing the four possible combinations of radiological diagnosis by either counting mixed diagnosis (+/– and –/+) as "overall negative" (left part) or "overall positive" (right part).

	++ = +	++,+–,–+– = +
	+–,–+,– – = –	– – = –
Odds Ratio	46	14
Threshold Value	0.05	18
Sensitivity	0.61	0.94
Specificity	0.97	0.47
Crude Error Rate	0.33	0.13

the fact that for most researchers a high sensitivity is preferred over a high specificity.

As an alternative, it can be concluded that there is a condition resembling LDH, which by both CT and myelography is lightly mistaken for LDH, but is not considered as "true LDH" during surgery. Candidates for this condition are scarring, bulging disk and spinal stenosis [56].

c.2. Two tests — reduction to a binary choice

The above mentioned fact that a combined CT and myelography investigation must result in a *yes or no decision* to operate the patient, implies that a choice has to be made between the two alternative methods of Table 2c. The first alternative is to considering two *negative* radiological tests and the mixture of a negative and positive test an overall *negative* test. The second alternative is to considering two *positive* radiological tests and the mixture of a positive and a negative test to be an overall *positive* test. From Table 2c can be easily seen that the first alternative is superior to the second in odds ratio. However, in practice the low sensitivity (0.61) and the high error rate (0.33) of the first alternative is considered unacceptable, so that alternative two is preferred [55]. This can be supported by the fact that none of the reports from the literature, reviewed in [36] has a TV of less than 0.8.

c.3. Two tests — order of performance

If two tests, i.c. myelography and CT, are sequentially performed, the order of performance does *not* influence the overall sensitivity and specificity. [*Note 11*] Also, if two tests are *conditionally* sequentially performed (i.e. the performance of the second test depends upon the results of the first one), the order of performance does *not* influence the overall sensitivity and specificity, provided that the decision rule is symmetric. In more concrete terms: if a patient with a negative CT or myelography (for LDH) at first investigation is submitted to the other procedure (myelography and CT, respectively), and a patient with a positive CT or a positive myelography at first investigation is not investigated further, it does not matter whether CT or myelography is performed first, if only the gain in information of the tests is considered.

In this case, patients with a positive first test are saved the burden, costs and risks of the second one. Therefore, from this point of view, it appears appropriate to start with the test that has the highest sensitivity (in our case: myelography, 82%, as compared with CT, 73%, table 2).

Other aspects of the test procedure, like whether it is an in- or out-patient procedure, the x-ray dose involved, whether it is an invasive procedure or not, a delay because of limited availability, costs (to the patient or society), etc., may be

reasons to start with one particular procedure. Most of these points are in favor of CT. All together it appears to be a good choice to start with CT, followed by myelography for those patients with a negative CT [55].

IV Complications

Most treatments in medicine carry the risk of unwelcome side-effects, as do some diagnostic procedures (invasive or radiologic ones, for instance). Occasionally, when the effectivities of two treatments are difficult to discriminate, a clear difference in the expected complications can indicate the best choice.

In the discussion about chemonucleolysis (CNL), the complications have drawn much attention and stirred up a vehement discussion. (See for example the gentlemen's dispute by Knox and Javid [42].) For a balanced judgement, complication rates need to be assessed for both CNL and surgery. Granted that it is difficult to assess a frequency for each complication, it is nevertheless remarkable that the literature does not even agree on a pure qualitative list of complications for these treatments [38]. This is more striking, because many patients appear to be very risk averse, regarding gambles with health-related outcomes [25].

a. Complications of CNL

The complications of CNL are of three kinds: allergic, neurologic and infectious [34, 35]. There are only a few sources in the literature which provide original data on a sufficiently large sample of patients [1, 46, 62]. The criteria for complications are difficult to standardize. This is especially critical for the allergic reactions, which can range from a mere hypotension to a full scale shock. In some studies ¼ - ⅓ of the patients are said to have severe spasms and pain immediately after CNL [3, 32, 46, 62], while most authors do not mention this at all. This discrepancy is most likely due to a different definition of complications [43]. Publications that give an enumeration of both minor and major complications [3, 32, 43], make it clear that there is probably a wide range and a gradual scale of adverse events that is related to this type of treatment. On the other hand, because of two factors, the complications are relatively well-known: (1) post-marketing surveillance by the industry, which was intensified for this product (chymopapain), ensured that complications were registered for a large number of patients [1], and (2) the fact that the major complication, allergic reaction, elicited research into causal mechanisms and preventive measures.

The findings from the literature can be summarized as follows (all numbers are per 1000 treatments) [34]

mortality:	about 0.5
anaphylactic reaction, including shock:	4-15
neurologic complications:	0.9-2.5
discitis:	0.8-1.6

b. Complications of LDH surgery

Complication assessments of LDH surgery as mentioned in the literature (for a summary see [38], table 6 (*p. 96 of this book*)) suffer from the same weaknesses as those of CNL. Only a difference in definitions and thresholds of complications is likely to explain the discrepancies between the high complication rates in two multicenter studies [49, 54] and the much lower rates in most of the other literature. The findings from the literature (since 1960) can be summarized as follows (again numbers per 1000 treatments):

mortality:	1-3
neurological damage (associated with surgical mishap):	2-10
infections, minor (wound –, urinary tract –):	10-100
infections, major (meningitis, osteomyelitis, discitis):	5-20

Relevant complication rates assessed from a discharge diagnosis survey (see below) are (per 1000 treatments):

mortality:	3
meningitis, encephalitis:	3.6
complications during surgery:	20 specified plus 10 unspecified.

c. Co-factors of complications

A problem with integrating crude complication rates from separate sources in the literature is that the co-factors of complications are generally not known. These may be *direct co-factors,* that are hard to assess, such as quality of the selection process of patients for surgery (advising conservative [*Note 12*] treatment for patients with contra-indications), but also the common *indirect co-factors,* such as age and sex, and *concurrent diseases* (or pathologic states), which are in general a risk for any surgical treatment.

These co-factors can only be accounted for if they are known *per patient* together with the complications. In an attempt to examine complication rates in a large sample (of USA hospital patients), with consideration of some of the above mentioned co-factors, complication rates were assessed for 3289 LDH surgery

patients and 4025 LDH conservatively treated patients drawn from the National Hospital Discharge Survey [38].

Age distributions appeared to be almost identical for LDH surgical and conservative patients, quite unlike Hudgins [31]; age appeared not to be significantly (at 5%) related with the complication rate. Concurrent diseases like hypertension, diabetes and clinical obesity appeared to be related with the risk of complications for surgery patients. Even though the numbers of patients experiencing complications associated with these risk factors are not very large, the associations are strong, compared with the findings in many epidemiological studies (and significant; O.R. approx 3-4, $p < 0.01$) [38].

d. Conclusions

It appears from the above that the complications for surgery are more important than for CNL, regarding both mortality and other complications.

This may lead to two conclusions of a different kind: (1) surgery should be more effective than CNL (and conservative treatment) to be an acceptable alternative, and (2) it is amazing that the complication rates of surgery have played a role *quantitatively* in the controversy around the safety of CNL only since 1986 [59].

V Outcome

There are many papers that report in one way or another on the results of treatment of LDH. Measures used include neurological signs [6, 32], physical examination (e.g. mobility of the spine) [3, 6], functionality (ADL) [9], ability to resume work [9, 32, 46, 64, 66], pain complaints [3, 15, 32, 46, 64, 66], use of analgesics [9, 32, 46, 64], subjective improvement [3, 9, 15, 32, 46], satisfaction with the treatment [15, 64], even whether patients would recommend the same treatment to their friends (which resulted in an interesting paradox, i.e. subjects would not recommend to their friends the treatment that they thought best) [66], and a considerable number of composite measures [9, 32, 64, 66]. All these measures represent a *different aspect* of the outcome of LDH treatment. Moreover, the exact wording of the questions on the same aspect can influence the answers considerably [30]. As it is important for a good decision to consider the above aspects not only in a period of 0.5-2 years, that most reports span, but also in the long run, it is apparent that there are insufficient data to give a *coherent* answer on all these aspects of outcome. Even if only the patient's subjective improvement is considered, there are only two papers that give repeated measurements over a longer period and present the data in a consequent way [22, 64]. Of these only one is a randomized trial, while the other is the only one that presents long term data for

operated patients with and without an abnormal disk. (This is important for the specificity of therapy.) Some global observations can be made from the literature:

a. First year after treatment
A possible favorable result of treatment is usually not present *at once* in all cases, but may come after a few months [43]. The success rates of most therapies for LDH increase during the first half year. This has been demonstrated for conservative therapy [22], surgery [22, 64], CNL [14, 15, 32, 43, 46], placebo injection [14, 32].

b. Long term
After one year further improvement is rather limited for *sciatica,* demonstrated for CNL and placebo [15], and for surgery and conservative therapy [22]. For CNL this is also supported by comparing many short term reports with long term reports, see [34, 35].

c. Speed of recovery
Patients having undergone surgery seem to improve somewhat faster than after conservative treatment [22], and CNL patients seem to improve faster compared with placebo patients [15].

d. In the end
After 7-10 years no clear difference can be found between surgical and conservative patients [22], nor between surgical and CNL patients [66]. This statement has to be treated with caution, as only limited data are available. [*Note 13*]

e. Relapses
Sciatica is not a complaint that is continuously present since there are remissions and relapses. This makes that any follow-up study that takes a snapshot of the patient's situation will probably not represent outcomes accurately. In the current analysis, the presence of relapses (2-4% per year [22], 3-4% per year on average [64] for 10 years) will only be used to point out that any report of treatment of LDH that claims a success rate of more than 90-95% should be carefully scrutinized for optimistic biases.

f. Later surgery
Another factor giving a clear indication of the upper limit of the effectiveness of treatment is whether "later surgery" is required, after a failed conservative, CNL, or surgical therapy. Twenty-six percent of the conservative treatments are followed by surgery in Weber's material [64], Javid indicated that 30-60% of CNL was followed by surgery (from 6 weeks to 6 months afterwards) [32], and Van Alphen reported 25% surgery within one year after CNL [3]. As it is a more "natural course" to proceed from conservative therapy or CNL to surgery, than to have a second operation, it would be misleading to compare the former rates with the

reoperation rate, which is 4% in the report of Smyth [58], and 3% in Van Alphen [3]. Hakelius reports that 9.3% of his operated patients had a disk operation before [22], and this number was approximately 7% with Gurdjian [21].

g. Placebo
There is a lower limit that can be posed upon the effectiveness of surgical treatment of LDH: that of a placebo treatment (the type of placebo should not make much difference). The absolute non-activity of a placebo is a very difficult question [19], but if we leave aside this issue, which played also a role in the controversy over the first CNL trials, the (placebo) effect of intradiscal injection of saline (55% [45], 47% [32], 47% [15]) is comparable to that of surgery on patients with normal disks: approximately 50% [31], 20-75% depending upon criterion [22]. (The percentages mentioned are "success rates", i.e. the fraction of all patients that state that the treatment was successful.)

h. Surgery without a herniation
Hakelius reports explicitly "no clear difference" in outcome for operated patients with or without an LDH found during surgery [22]. This is potentially a very important conclusion, as it would decrease the need for an exact diagnosis of LDH before operation. His conclusion is not confirmed by the other reports mentioned under g. There is always the problem of small numbers of this category of patients. [*Note 14*]

i. The position of CNL
From the randomized trials, CNL appears to be superior to a placebo in terms of effectiveness, but inferior to surgery. A more precise conclusion is hardly justifiable.

j. Regression and recurrence of an LDH
It is possible for a herniation to disappear "spontaneously", i.e. without conservative treatment [60]. This should be kept in mind if discrepancies between clinical signs, radiology and surgical findings are considered, and also influences our view on the natural course of the disease. LDH's can also recur, either at the same level and side, or at another localization [21, 50, 58]. This also helps to explain the erratic course of sciatica in some patients.

k. Conclusion from the observations

Combining these observations (*a.* —*j.*), it appears that the "natural course" of an LDH (measured by placebo and conservative treatment) is in the long run (at least one year) not distinguishable from the course after CNL or surgery. The latter two treatments may provide faster relief during the first year, and — in the case of surgery — may prevent part of the relapses that would otherwise have occurred.

Consequently, the main gain of surgery and CNL is in the first year. The one year effectiveness of *any* treatment is limited at the lower limit by the placebo effect (approximately 50% success) and at the upper end by residual sciatica, low back pain and relapses, and will consequently likely not exceed 90%. Given the multidimensional nature of the outcome and the large inaccuracy of the measurement of the patient's subjective improvement, it is clear that it will be hard to establish the superiority of one treatment over the other in the span that is left open between these two limits. The multitude of rather confusing data indicates that the above mentioned order of effectiveness (surgery better than CNL, which is better than placebo) is the most likely. It is harder to position conservative treatment; it appears to be initially poorer than surgery, but is it better than, equal to, or worse than CNL? The answer is a matter of belief, which is a widely used psychopharmacon and adjuvant therapy in sciatica.

Let us assume for a moment that we are fortunate to have the answer from an ideal study, and see what it would teach us. If we would know, using unbiased and accurate measurements and extremely large samples, that, e.g., the one-year "overall effectiveness" of placebo would be exactly 50%, of conservative therapy exactly 65%, of CNL exactly 70% and of surgery 85%, while the common end results would be 80%, these data by themselves would still be insufficient for us to make a choice. (The percentages are, again, success rates.) In order to answer the question of which treatment to choose, the complications of all three treatments must be considered. Further, each patient has to consider for himself how much he can endure and wait, and how important to him would be a doubling of the odds of favorable results in the first year, compared with the additional risks and the inconvenience of surgery.

VI Utilities

In the three previous sections the need for *utilities* was illustrated. [*Note 15*] Utility is the value which a person or a party attributes to a specified (perhaps future or hypothetical) event, and is usually expressed in relative units, as it is much easier to compare the utilities of two events, than to give the absolute utility of a single event. Utilities do not play an explicit role in the major part of the clinical literature. Yet they are ubiquitous in everyday clinical practice.

a. Utilities related to investigations

In the analysis of the relative diagnostic qualities of CT, myelography, and phlebography (section III), a non-negligible [*Note 16*] threshold value component has been demonstrated. Of course this could be caused by the particular mix of

radiological discriminating features within each of the procedures CT, myelography or phlebography, but it is much more plausible that it is caused by preferences of the observers for specific or sensitive answers. This view can be supported by three arguments:

(1) the fact that the typical and extreme values of specificity and sensitivity are common for all three radiological procedures;
(2) the experience that clinicians, including radiologists, sometimes make statements about a preference for sensitive or specific results; [*Note 17*]
(3) the fact that the *inevitability* of a preference for specific or sensitive results can be demonstrated for the special case of a combination of two dichotomized tests (CT and myelography, see section III).

There may be a good reason for the physician to use his own utilities in giving a final diagnosis: the radiological image often serves to support the hypothesis that a patient has an LDH and does not serve to support the hypothesis that the patient's LDH-like complaints are caused by something else. Perhaps the patient even profits from this situation, because it has yet to be proven that the decisions where the patient gives his own utilities are in the long run better than when the physician has some influence on the patient's utilities. From the point of view of the discipline of *clinical decision analysis,* where a patient is supposed to supply his own utilities, it would be preferable for physicians to be neutral with respect to false positive and false negative errors (and risk neutral as well), or, alternatively, to tell a patient besides the probabilities that he has an LDH (or not) the probabilities that these (first) probabilities are incorrect. [*Note 18*]

b. Utilities of complications

Assessing utilities for complications is usually more difficult than for treatment results for several reasons:

(1) complications are rare (for most diseases and treatments), making it harder for most patients to give consistent utilities for situations that are hard to imagine and accept. This implies that utilities, in the human mind, are hard to separate from probabilities [57].
(2) complications are more diverse, than treatment results, ranging from wound infection to death, with meningitis and shock in between.
(3) most patients do not want to hear about every possible complication that can happen. ("The negative utility of assessing utilities" [28]) The list of what can possibly go wrong is quite large. [*Note 19*]

For the particular set of complications that emerges from the literature on the surgical and chirurgico-chemical treatment of LDH, no (formal) utility assessment has ever been made, to our knowledge. The best that can be done at present — without a major effort in data collection (with tricky ramifications in the realm of psychology [33], and in the field of generalizing utilities above the individual level

[27]) — is to compare complications in a few distinct categories (see section IV). Fortunately, in our case, all categories appear to have more complications following surgery than following chemonucleolysis. (*probability dominance* [40] [*Note 20*]). Regardless of the utilities involved, only one conclusion concerning complications seems justified: chemonucleolysis appears to be associated with complications to a lesser degree than surgery.

c. Utilities of outcome

The multi-faceted nature of therapeutic outcomes (see section V) makes utility assessment of results quite complicated. Many aspects of outcome are not applicable to every patient, a detail that is not yet incorporated into most of the utility-analytic methods. A few simplifications in the description of the treatment outcome can reduce the utility assessment to the question of whether a patient is willing to accept the risk of surgical treatment in order to improve — on average — faster. In clinical practice this question is often posed to patients in the doctor's office — in non-quantitative terms.

More difficult is whether a patient prefers a less risky, but less (immediately) effective treatment (CNL) over surgery. Based on the precision of the current estimates of treatment results, the above plain question to the patient may likely have more meaning than a formal decision analysis.

d. Theoretical considerations

Ranking utilities directly by the patient on a scale *à l'improviste* is seldom possible. Therefore (on an individual level) relative utilities are usually assessed by having a person (patient) answer questions on his preference for varying outcomes or varying probabilities of fixed outcomes. In these so-called "hypothetical gambling" questions, e.g. an indifference gamble or a equivalence gamble (see [61] and chapter 6 of [65]), the person (patient) is supposed to be fairly consistent. The ways in which he is allowed not to stick to the arithmetic equivalence [*Note 21*] of an expected utility are fairly restricted. Some recognized ways in which persons (patients) are allowed to be "consistently irrational" [*Note 22*] are *risk proneness/aversion,* and *non-linearity* of their utility with the quantity of the expected outcome or with the probability of an outcome (as long as monotonicity is maintained) [40]. Most of the literature does *not* deal with the situation that a person who is risk prone at a certain moment or for a certain event, may be risk averse at another moment or for another event. Yet this situation is part of the human condition and is common place in clinical practice. It does not help to say that such a person is behaving irrationally.

Applications of such utility assessment (i.e. with hypothetical gambling questions) exist in problems with outcomes which have several aspects, as is the case with LDH treatment. By simplifying matters, it is possible to assess utilities of partial aspects, provided these are mutually independent. [*Note 23*] A classical example is the life expectancy and the quality of life during the remaining life years. The multi-attribute utility approach can cope in principle with outcomes which have many aspects, but imposes additional requirements [*Note 24*] on the analytic power of the patient who is to give his personal values and thus carry with him the responsibility for the decision for the rest of his (correspondingly quality adjusted) life.

The applicability of the above constraints on assessment of utility may give the reader some concern about the feasibility of utility assessment in the clinical practice of LDH treatment. It is hoped that more detailed studies of the course of treatment results, together with the developments in practical utility assessment will make such possible, an agenda which reaches far beyond the scope of this dissertation.

Notes

1 More precisely, new information *can* be gained, if there is a fluctuation in the complaints and symptoms. In our case, the history which the patient is presenting to the physician makes that not only a snapshot is taken from the course of the disease, but that a period of some duration is considered.

2 All three simplifications do not hold completely, obviously. Models incorporating several corrections for these "biases" can be formulated, but even moderately realistic models will become too complicated. Moreover, there is a lack of data on which to fit the models. Therefore, the simplest possible approach is used here.

3 For the calculations and intepretations of the odds ratio in different contexts, the reader is referred to [37], [18] and [41].

4 The meaning of false positive and false negative errors is clear;it is formed by all patients with a positive test and no disease, and all patients with a negative test and the disease respectively. About the related measures, false positive rate and false negative rate, there is considerable confusion. Hence, the definition of these terms in this report is presented here: FNR is $P(T-|D+)$, and FPR is $P(T+|D-)$, in accordance with [65].

5 For the notation, calculation and interpretation of the threshold value see the appendix of [37] (*p. 63 in this book).*

6 It may be considered a corroboration for the fact that this TV range is more determined by the observer than by the radiogical test, that the same typical and extreme values hold for phlebography as well.

7 That this is not impossible can be derived from the fact that the postulated OR of 60 is taken from the estimated OR of phlebography (see [37]). Several decades and

a good number of publications were required, however, to make this assessment possible.

8 Nearly always, that is: if $0 < P(LDH|MY+) < 1$ and $0 < P(LDH|CT+) < 1$, and if $P(LDH|MY+) \neq P(LDH|MY-)$ and $P(LDH|CT+) \neq P(LDH|CT-)$.

9 Other potential sources are [7], [24], and [48]. The first did not contain a *complete* 2×2×2 table. The second and third had limited numbers of cases, resulting in empty cells, or with very low numbers. This would make a sub-analysis not reliable.

10 There is a dire need for more of these in the literature. It is still open whether the same interaction as described above will be found by other observers.

11 The following conditions are supposed to hold: the test results do not influence each other; the test results do not influence the patient. An example of a test influence on the patient is the contrast material that is present for some time after a myelography.

12 The term conservative treatment is used in this publication because it is the common term for non-surgical treatment, mainly consisting of bedrest. No implication that invasive treatment cannot be conservative in the original sense of the word is to be assumed.

13 The material of Weber [63, 64] deserves special attention: his patients still showed improvement one year after surgery or conservative treatment (b). Yet, if another (slower) time scale is applied, his observations support our conclusions under (c) and (d).

14 Espersen [13] provides interesting data that do not fit in our scheme, however. In 134 patients operated on clinical grounds only, 41% had normal disks. From sixteen patients operated while having normal myelograms, 50% had a good outcome of surgery.

15 Some sources make a distinction between "value", a psychological, subjective, entity and "utility", an objective entity; in this text the words utility and value are used interchangeably.

16 The exact magnitude is hard to establish, because of the several biases in the approach and some unknown parameters.

17 Some examples: Herkowitz e.a. [26] write: "The patients [...had...] a positive tension sign and/or neurological deficit. [...] In addition a corroborating positive contrast study was necessary before surgical intervention was undertaken."
Anand e.a. [4] only used phrases like "(failed to) demonstrate an LDH".
Gulati e.a. [20] in a reasoning that is quite close to ours (Schipper e.a. [55]) write: "If indeterminate or indifferent clinical signs of an HNP are present, and a [...] CT scan shows a normal disc space [...] myelography need not to be performed. If the clinical findings are definite [...] myelography [...] should be performed."

18 This is not meant as a proposal for a practically feasible solution, as many patients cannot be expected to be able to use these four probabilities correctly, but physicians can. Not all of the physicians currently do this, however [12].

19 An illustration may be that the warnings on the instructions for use of pharmaceutical products falsely gives many lay people the impression that acetylsalicylate is a quite dangerous drug.
20 There exist some other mechanisms by which matters can be simplified and a satisfying utility scale can be constructed. One of these is that a *sensitivity analysis* shows that any remaining imprecision in utility scaling does not matter.
21 E.g. the event of a loss of $100 with probability of 0.5 is equivalent to the event of a loss of $200 with probability 0.25 or a certain loss of $50.
22 Sometimes even formulated as that the decision maker is required to be "irrationally rational", i.e. rational beyond what is reasonable.
23 There are partial solutions for the case where these are not independent. (See [51], section 4.5.7)
24 Besides the scaling factors for the different attributes that have to be supplied, it is clear that the patient has to give more (for him unusual) parameters in a consistent way, which is in itself a difficult task.

References

1 Agre K, Wilson RR, Brim M, McDermott DJ: Chymodiactin post-marketing surveillance — demographic and adverse experience data in 29,075 patients. *Spine* 1984; **9**: 479-485.
2 Aitken AP, Bradford CH: End results of ruptured interverbtebral discs in industry. *Amer J Surg* 1947; **73**: 365-376
3 Alphen HAM van, Braakman R, Bezemer PD, Broere G, Berfelo MW: Chemonucleolysis versus discectomy: a randomized multicenter trial. *J Neurosurg* 1989; **70**: 869-875.
4 Anand AK, Lee BCP: Plain and metrizamide CT of lumbar disk disease: comparison with myelography. *AJNR* 1982; **3**: 567-571.
5 Bell GR, Rothman RH, Booth RE, Cuckler JM, Garfin S, Herkowitz H, Simeone FA, Dolinskas C, Han SS: A study of computer assisted tomography. II. Comparison of metrizamide myelography and computed tomography in the diagnosis of herniated disk and spinal stenosis. *Spine* 1984; **9**: 552-556.
6 Blaauw G, Braakman R, Gelpke GJ, Singh R: Changes in radicular function following low-back surgery. *J Neurosurg* 1988; **69**: 649-652
7 Bosacco SJ, Berman AT, Garbarino JL, Teplick JG, Peyster R: A comparison of CT scanning and myelography in the diagnosis of lumbar disc herniation. *Clin Orthop* 1984; **190**: 124-28.
8 Coin CG, Chan YS, Keranen V, et al.: Computer assisted myelography in disk disease. *J Comput Assist Tomogr* 1977; **1**: 398-404.
9 Crawshaw C, Frazer AM, Merriam WF, Mulholland RC, Webb JK: A comparison of surgery and chemonucleolysis in the treatment of sciatica. *Spine* 1984; **9**: 195-198.

10 Deburge A, Rocolle J, Benoist M: Surgical findings and results of surgery after failure of chemonucleolysis. *Spine* 1985; **10**: 812-815.
11 Di Chiro, G: Computed tomography of the spinal cord after lumbar intrathecal introduction of metrizamide (computer-assisted myelography). *Radiology* 1976; **120**: 101-104.
12 Eddy DM: Probabilistic reasoning in clinical medicine: problems and opportunities. p 247-267 in D Kahneman, P Slovic, A Tversky, *Judgement under uncertainty, heuristics and biases.* Cambrige, Cambridge University Press, 1982.
13 Espersen JO, Kosteljanetz M, Halaburt H, Miletic T: Predictive value of radiculography in patients with lumbago-sciatica - A prospective study (part II). *Acta Neurochirurgica* 1984; **73**: 213-221.
14 Feldman J, Menkes CJ, Pallardy G, Chevrot A, Horreard P, Zenny JC, Godefroy D, Amor B: Etude en double-aveugle du traitement de la lombosciatique discale par chimionucléolyse. *Revue du Rhumatisme* 1986; **53**: 147-152.
15 Fraser RD: Chymopapain for the treatment of intervertebral disc herniation - The final report of a double-blind study. *Spine* 1984; **9**: 815-818.
16 Friedman WA: Percutaneous discectomy: an alternative to chemonucleolysis? *Neurosurgery* 1983; **13**: 542-547.
17 Fries JW, Abodeely DA, Vijungco JG, Yeager VL, Gaffey WR: Computed tomography of herniated and extruded nucleus pulposus. *J Comput Assist Tomogr* 1982; **6**: 874-887.
18 Green DM, Swets JA: *Signal Detection Theory and Psychophysics.* New York, John Wiley & Sons, Inc, 1966. (Reprinted in 1974 by R.E. Krieger Pub Co, Huntington, NY.)
19 Gross F: Placebo - the universal drug. *Meth Inf Med* 1984; **23**: 176-182.
20 Gulati AN, Weinstein Z, Studdard W: CT scan of the spine for herniated discs. *Neuroradiology* 1981; **22**: 57-60.
21 Gurdjian ES, Ostrowksi AZ, Hardy WG, Lindner DW, Thomas LM: Results of operative treatment of protruded and ruptured lumbar discs based on 1176 operative cases with 82 percent follow-up of 3 to 13 years. *J Neurosurg* 1961; **18**: 783-791.
22 Hakelius A: Prognosis in sciatica. A clinical follow-up study of surgical and non-surgical treatment. *Acta Orthop Scand* 1970, Suppl. 129.
23 Hampton AO, Robinson JM: Roentgenographic demonstration of rupture of the intervertebral disk into the spinal canal after injection of Lipiodol. *AJR* 1936; **36**: 782-803.
24 Haughton VM, Eldevik OP, Magnaes B, Amundsen P: A Prospective comparison of computed tomography and myelography in the diagnosis of herniated lumbar disks. *Radiology* 1982; **142**: 103-110.
25 Hellinger FJ: Expected utility theory and risky choices with health outcomes. *Medical Care* 1989; **27**: 273-279.
26 Herkowitz HN, Wiesel SW, Booth RE, Rothman RH: Metrizamide myelography and epidural venography. *Spine* 1982; **7**: 55-64.
27 Hilden J: The non-existence of interpersonal utility scales. A missing link in medical utility theory. *Med Decis making* 1985; **5**: 215-228.

28 Hilden J: Utility analysis. *Theor Surg* 1987; **2**: 133-137.
29 Houtteville JP, Toumi K: Constatations opératoires et resultats après échec de la chimionucléolyse lombaire. *Neurochirurgie* 1985; **31**: 494-498.
30 Howe J, Frymoyer JW: The effects of questionnaire design on the determination of end results in lumbar spinal surgery. *Spine* 1985; **10**: 804-805.
31 Hudgins WR: The predictive value of myelography in the diagnosis of ruptured lumbar discs. *J Neurosurg* 1970; **32**: 152-162.
32 Javid MJ, Nordby EJ, Ford LT, Hejna WJ, Whisler WW, Burton C, Millet K, Wiltse LL, Widell EH, Boyd RJ, Newton SE, Thisted R: Safety and efficay of chymopapain (chymodiactin) in herniated nucleus pulposus with sciatica. *JAMA* 1983; **249**: 2489-2494.
33 Kahneman D, Slovic P, Tverksy A: *Judgement under uncertainty: heuristics and biases.* Cambridge, Cambridge University Press, 1982.
34 Kardaun JWPF: *Effectiviteit en bijwerkingen van chemonucleolyse.* [Achtergrondstudie nr 1987-19]. Den Haag, Gezondheidsraad, 1987.
35 Kardaun JWPF, Schipper J: Chemonucleolyse — effectiviteit en bijwerkingen. *Ned Tijdschr Geneeskd* 1988; **132**: 285-289 *Chapter 4 in this book.*
36 Kardaun JWPF, J Schipper, R Braakman: CT, myelography and myelography in the detection of lumbar herniated disk — Review of the literature. *Am J Neuroradiology* 1989; 10: 1111-1122. *Chapter 2 in this book.*
37 Kardaun JWPF, OJWF Kardaun: Test qualities of three radiological procedures for the detection of lumbar disk herniation. In press (a), *Information in Medicine.* *Chapter 3 in this book.*
38 Kardaun JWPF, LR White, WO Shaffer: Acute complications in patients with surgical treatment of lumbar herniated disk. In press (b), *Journal of Spinal Disorders.* *Chapter 5 in this book.*
39 Kassirer JP, Moskowitz AJ, Lau J, Pauker SG: Decision Analysis: A progress report. *Ann Int medicine* 1987; **106**: 275-291.
40 Keeney RL, Raiffa H: *Decisions with multiple objectives — preferences and value tradeoffs.* New York, John Wiley & Sons, 1976.
41 Kleinbaum DG, Kupper LL, Morgenstern H: *Epidemiologic Research — Principles and quantitative methods.* Belmont, CA, Lifetime Learning Publications, 1982.
42 Knox DL: Efficacy of chymopapain chemonucleolysis. [letter] [+ reply MJ Javid] *J Neurosurg* 1986; **64**: 162-163.
43 Leeuwen RB van: *Chemonucleolysis — Beloop, resultaten en prognostische factoren. Ph.D. thesis, Utrecht, 1989.*
44 Maroon JC, Abla A: Microdiscectomy versus chemonucleolysis. *Neurosurgery* 1985; **16**: 644-649.
45 Martins AN, Ramirez A, Johnston J, Schwetschenau PR: Double blind evaluation of chemonucleolysis for herniated lumbar discs: late results. *J Neurosurg* 1978; **49**: 816-827.
46 McDermott DJ, Agre K, Brim M, Demma FJ, Nelson J, Wilson RR: Chymodiactin in patients with herniated lumbar intervertebral disc(s) - An open label, multicenter study. *Spine* 1985; **10**: 242-249.

47 Modic MT, Masaryk T, Boumphrey F, Goormastic M, Bell G: Lumbar herniated disk disease and canal stenosis: prospective evaluation by surface coil MR, CT, and myelography. *AJR* 1986; **147**: 757-65.
48 Moufarrij NA, Hardy Jr RW, Weinstein MA: Computed tomographic, myelographic and operative findings in patients with suspected herniated lumbar discs. *Neurosurgery* 1983 **12**: 184-88.
49 Oppel F, Schramm J, Schirmer M, Zeitner M: Results and complicated course after surgery for lumbar disc herniation. *Adv Neurosurg* 1977; **4**: 36-51.
50 Pheasant HC: Sources of failures in laminectomies. *Orthop Clinics Nth America* 1975; **6**:1,319
51 Raiffa H: *Decision analysis.* Reading, MA, Addison-Wesley, 1968.
52 Saal JA, Saal JS: Nonoperative treatment of herniated lumbar intervertebral disc with radiculopathy; an outcome study. *Spine* 1989; **14**: 431-437.
53 Sachsenheimer W, Hamer J, Müller HA: The value of spinal computed tomography in diagnosis of herniated lumbar discs. *Acta Neurochirurgica* 1982; **60**: 107-114.
54 Schepelmann F, Greiner L, Pia HW: Complications following operation of herniated lumbar discs. *Advances in Neurosurgery* 1977; **4**: 53-54.
55 Schipper J, JWPF Kardaun, R Braakman, KJ van Dongen, G Blaauw: Lumbar disk herniation: Diagnosis with CT or myelography. *Radiology* 1987; **165**: 227-231. *Chapter 1 in this book.*
56 Slebus FG, Braakman R, Schipper J, Van Dongen KJ, Westendorp-de Serière M: Non-corresponding radiological and surgical diagnosis in patients operated for sciatica. *Acta Neurochirurgica (Wien)* 1988; **94**: 137-143.
57 Slovic P, B Fischoff, S Lichtenstein: Facts versus fears: understanding perceived risk. P 464-489 in D Kahneman, P Slovic, A Tversky, *Judgement under uncertainty, heuristics and biases.* Cambrige, Cambridge University Press, 1982.
58 Smyth H, Gallagher J, McManus F: Surgery in lumbar disc protrusion - a long term follow-up. *Ir Med J* 1983; **76**: 25-26.
59 Sutton JC: Chemonucleolysis — current status and future outlook. *Neurochirurgia* 1986 (a); **29**: 173-178.
60 Teplick JG, Haskin ME: Spontaneous regression of herniated nucleus pulposus. *AJNR* 1985; **6**: 331-335.
61 Vlek CAJ, Wagenaar WA: Oordelen en belissen in onzekerheid. P 447-492 in: JA Michon, EGJ Eijkman, LFW de Klerk, *Handboek der Psychomomie.* Deventer, Van Loghum Slaterus 1976.
62 Watts C: Complications of chemonucleolysis for lumbar disc disease. *Neurosurgery* 1977; **1**: 2-5.
63 Weber H: Lumbar disc herniation. *J Oslo City Hosp* 1978; **28**: 33-64, 89-120.
64 Weber H: Lumbar disc herniation. A controlled, prospective study with ten years of observation. *Spine* 1983; **8**: 131-140.
65 Weinstein MC, Fineberg HV, Elstein AS, Frazier HS, Neuhauser D, Neutra RR, McNeil BJ: *Clinical decision analysis.* WB Saunders Co., Philadelphia 1980.

66 Weinstein J, Spraat KF, Lehmann T, McNeill T, Hejna,W: Lumbar disc herniation. A comparison of the results of chemonucleolysis and open discectomy after ten years. *J Bone Joint Surg* 1986; **68A**: 43-54.

SUMMARY

The diagnosis and treatment of LDH have the attention of many clinicians and researchers. The present study attempts to shed new light on the matter, not primarily by presenting new facts, but largely by rearranging and reconsidering information already available.

The focus is not on which clinical sign is the best predictor of LDH, nor on which aspect of myelography is the best criterion for LDH, but rather on *how much radiology can add* to the clinical diagnosis, whether *one radiological procedure is sufficient,* (and should this be CT or myelography), or whether we need two. Likewise, regarding treatment the focus is not on which variant of surgical technique is best but on *how effective and safe* treatment is. In aiming at these more global questions, some simplifications had to be carried out and some interesting details had to be neglected.

Only part of the trajectory that is traversed by the patient with a lumbar radicular syndrome due to an LDH is considered in this study. The patient is supposed to be seen by the general practitioner and possibly a neurologist before he arrives at the (neuro)surgeon's office. The clinical diagnosis is usually fortified by radiological investigation. At present, the two most readily considered procedures are CT and/or myelography.

Using data from a large part of the relevant literature, a distinction was made in test-related differences (odds ratio) and observer-related differences (threshold value). Myelography appeared to give slightly better results than CT (not considering secondary factors such as cost and risks). There is always a time lag between the use of published data and the state of the art. Therefore, it should be kept in mind that this rank order may change.

A tendency can be seen for radiological LDH reports to be more sensitive than specific. Inter-observer differences appeared to be quite influential for the results of individual studies. This should probably be an integral part of reporting research on the quality of radiological procedures. The pattern of observer differences was not substantially different between CT and myelography.

When an LDH is confirmed by radiology, the three most prevailing treatments are conservative treatment, chemonucleolysis (CNL) and surgical treatment (discectomy). The most widely praised method for comparing two treatment modalities is the randomized clinical trial (RCT). RCT's for surgical treatment vs CNL

did show better results for surgery, while CNL compared favorably with placebo. These RCT's usually involved only a very limited number of patients, and were generally limited to a short follow-up period of about (half) a year. To overcome the narrow focus of these RCT's, other literature on the effectiveness of LDH treatment must be considered, thereby sacrificing some methodological purity. The pattern of results can, with some caution, be interpreted as being in agreement with the indicated rank order (surgery better than CNL, which is better than placebo). However it appears that after a few years the treatment results are hard to discriminate and are generally favorable for each of the three treatments considered. This would indicate that the main asset of surgical treatment would be a faster recovery. Put in these simple words, the patient is likely quite able to give his own personal preference (utilities) as to whether he prefers to wait and suffer in the mean time, or to accept the additional risks of surgery.

The complications of the various types of treatment for LDH have received unbalanced attention in the past. For *chemonucleolysis,* the risks have been quite extensively assessed at the time that this treatment was discredited and was temporarily withdrawn from the US market. The main concern was the possibility of an allergic reaction, for which a skin test has subsequently been developed. Even when this allergic reaction could be completely prevented (which is not the case) at the cost of withholding a small fraction of the patients the possibility of CNL, the other complications still would deserve attention. It is hard to say how serious these other complications are in absolute terms. Compared with the complications that can emerge from surgical treatment of LDH, however, it is clear that *surgical treatment* is associated with more complications in each of a few classes of severity. There is no equivalent for a skin test to predict complications in surgical LDH patients, but, based on an analysis of hospital discharge records, it seems prudent to be extra cautious in advising surgery for patients with general risk factors such as obesity, hypertension or diabetes.

Given the current imprecision in the assessment of treatment outcome and complications, a formal utility analysis in patients with LDH for the different options is not likely to be of practical value.

SAMENVATTING

Diagnose en behandeling van een lumbale hernia nuclei pulposi (LHNP) zijn een onderwerp van aandacht voor vele clinici en onderzoekers. Deze verhandeling poogt nieuw licht te werpen op dit onderwerp, niet in de eerste plaats door nieuwe feiten aan te dragen, maar door de bestaande feiten opnieuw, na herschikking, te presenteren.

De aandacht gaat niet uit naar welk symptoom de beste aanwijzing zou geven voor het bestaan van een LHNP of naar wat het beste myelografische criterium is voor een LHNP, maar naar vragen als hoeveel de radiologie kan bijdragen aan de klinische diagnose, of één radiologisch onderzoek voldoende is of dat er een tweede nodig is en welke methode — CT of myelography — het meest geschikt is. Evenzo ligt bij de behandeling de nadruk niet zozeer op welke variant onder de operatieve technieken het beste is, maar op de vraag hoe effectief de behandeling is en hoe veilig. Bij de beantwoording van deze globale vragen, moeten enkele vereenvoudigingen op de koop toe worden genomen en moeten enkele detailaspecten helaas onderbelicht blijven.

Slechts een gedeelte van de route die menig patiënt met een lumbaal radiculair syndroom doorloopt wordt in deze verhandeling in beschouwing genomen. De patiënt wordt geacht reeds door huisarts en eventueel neuroloog "gezien" te zijn. Vervolgens wordt hij verwezen naar neurochirurg of orthopeed. De klinische diagnose wordt meestal ondersteund met een radiologisch onderzoek, waarvan tegenwoordig CT en/of myelografie het meest in aanmerking komen.

Met behulp van gegevens, verkregen uit een groot gedeelte van de relevante literatuur, wordt een onderscheid gemaakt tussen onderzoek-gebonden verschillen (*odds ratio*) en beoordelaar-gebonden verschillen (*threshold value*). Myelography blijkt iets beter te zijn dan CT, als factoren als kosten en complicaties buiten beschouwing gelaten worden. Daarbij moet aangetekend worden, dat deze volgorde in de eerstkomende jaren kan veranderen. Hergebruik van gepubliceerde gegevens (*meta-analyse*) loopt noodzakelijkerwijze iets achter op de *state of the art*.

Uit deze gegevens blijkt een tendens dat radiologische diagnosen betreffende een LHNP eerder naar het sensitieve dan naar het specifieke neigen. De verschillen tussen de beoordelaars, de threshold values, bepalen in aanzienlijke mate de resultaten in de afzonderlijke publikaties, en dienen daarom een integraal onderdeel te worden van de rapportage van onderzoeksresultaten over de kwaliteit van

radiologische procedures. CT en myelografie tonen hetzelfde patroon van beoordelaar-gebonden verschillen.

Nadat een hernia door de radioloog bevestigd is, komen drie categorieën behandelingen in aanmerking: conservatieve (d.w.z. niet-operatieve) behandeling, chemonucleolyse (CNL) en operatieve behandeling (discusuitruiming). Voor het vergelijken van verschillende behandelingsvormen wordt bij voorkeur gebruik gemaakt van een *randomized clinical trial* (RCT). De RCT's die operatie met CNL vergeleken, toonden betere resultaten met operatie aan, terwijl CNL gunstig afstak in een RCT tegen een placebo. De betreffende RCT's onderzochten in de regel een nogal beperkt aantal patiënten en beperkten de follow-up tot een periode van een half jaar tot een jaar. Om een bredere kijk op de vergelijking te krijgen dan de RCT's toelaten, dient men zijn toevlucht te nemen tot de overige literatuur over de effectiviteit van behandelingen voor LHNP, hoewel dit ten koste gaat van de methodologische zuiverheid.

Het geheel der onderzoeksresultaten kan onder enig voorbehoud geduid worden dat het bovengenoemde volgorde ondersteunt (d.w.z. operatie is beter dan CNL, hetgeen op zijn beurt beter is dan een placebobehandeling). Een paar jaren na afloop van de therapie blijken er echter ternauwernood verschillen waarneembaar tussen de resultaten van deze vormen van therapie. De lange termijn resultaten zijn over het algemeen vrij gunstig voor elk van de drie beschouwde behandelingswijzen.

Deze constateringen leiden tot de conclusie dat het voornaamste voordeel van operatie gelegen zou zijn in een sneller, niet in een beter herstel. Indien de keuze in dergelijke eenvoudige termen omschreven wordt, zal het merendeel der patiënten best in staat zijn een eigen voorkeur aan te geven, en af te wegen of zij liever al lijdend afwachten, dan wel de risico's van een operatie aanvaarden.

De complicaties van de behandeling van een HNP hebben op ongelijke wijze, namelijk afhankelijk van de behandeling, aandacht gekregen. De risico's van *chemonucleolyse* zijn vrij uitgebreid in kaart gebracht toen het middel aan sterke kritiek blootstond en in de VS zijn registratie verloor. De belangrijkste reden was de mogelijkheid van een allergische reactie, die aanleiding was om een huidtest te ontwikkelen ter preventie hiervan. Maar zelfs indien hiermee een allergische reactie geheel voorkomen zou kunnen worden — hetgeen helaas niet het geval is — zijn de overige complicaties niet verwaarloosbaar gering. Een absolute maat voor de ernst daarvan is nauwelijks voorhanden. Bij een vergelijking met complicaties die uit *operatie* voort kunnen vloeien, is het echter duidelijk dat CNL in elk van een paar categorieën van complicaties gunstiger voor de dag komt. Een test om chirurgische complicaties te voorspellen — analoog aan de huidtest voor een overgevoeligheidsreactie voor chymopapaïne — bestaat niet; wel lijkt het op grond

van een analyse van gegevens uit ziekenhuis-ontslagregistraties, verstandig om iets terughoudender te zijn om operatie aan te raden aan patiënten met algemene risicofactoren, zoals overgewicht, hoge bloeddruk of diabetes.

Bij de huidige onnauwkeurigheid in het bepalen van het resultaat van de behandelingen en de aard en omvang van de complicaties, lijkt het verrichten van een formele utilitieitsanalyse bij patiënten met een LHNP van weinig praktisch nut te zijn voor het maken van een behandelingskeuze.

Curriculum Vitae

Jan Kardaun, born in The Hague in 1955, attended primary and secondary schools (gymnasium β) in Venlo, Raalte and Zwolle. In 1973 he started studying the arts and sciences of medicine at Groningen University. He developed an interest in what was called at that time "computer applications in medical care". After his internships at the Deventer Hospitals, and his medical graduation (1981) he turned towards a research career and found an opportunity at the Department of Public Health and Social Medicine, Erasmus University. During 1982 — 1987 he was engaged as *universitair docent* at this department in applied research of formal methods in medicine. In 1987 he obtained his certification as *sociaal geneeskundige / epidemioloog* (supervisor: Prof. Dr. P.J. van der Maas). In 1988 he was Guest Researcher at the National Institutes of Health. Since 1989 he is working as Chief of the Public Health Section, Department of Health Statistics, Netherlands Central Bureau of Statistics.