From the Department of Clinical Sciences,
Family Medicine Stockholm
Karolinska Institutet, Stockholm, Sweden

DIAGNOSING HEART FAILURE IN PRIMARY HEALTH CARE

Ylva Skånér

Stockholm 2004

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden

ISBN 91-7349-784-3

Livet är kort,
konsten lång,
rätta tillfället går snart förbi,
erfarenheten är bedräglig,
omdömet svårt.

Life is short, and Art long;
the crisis fleeting;
experience perilous,
and decision difficult.

Hippocrates

Bedömning

Man ser inte genast skillnad
på småsten och gråsparvar.

Några sätter sig i nypontörnet
- de är gråsparvar.

Andra blir kvar och trippar i åkern
- de är också gråsparvar.

Andra återigen ligger stilla kvar i åkern
- de är troligen stenar.

Anna Rydstedt 1976

ABSTRACT

Diagnosing chronic heart failure (CHF) is difficult. General practitioners (GPs) have an important role in the management of heart failure patients, and the purpose of the studies was to examine their judgements of patients with suspected CHF. Two methods from cognitive psychology were used, Clinical Judgement Analysis (CJA) in Studies I-IV, and think-aloud technique in Study V. Written case vignettes based on authentic patients were presented either in a paper format (Studies I-III) or on a computer screen (Study V). In Study IV, theoretical and practical problems concerning how to construct suitable case vignettes for CJA studies were discussed, with reference to experiences from Studies I-II.

In Study I, 27 GPs assessed the probability of CHF for 45 case vignettes, five of which were duplicates. Each GP's diagnostic strategy was defined as the set of statistical regression weights for the different variables (cues) describing the patient. Both judgements and strategies varied widely among the GPs, but according to analysis of the duplicate cases, the GPs were consistent in their judgements. The most important cues were pulmonary congestion and cardiac volume. In Study II, 27 GPs, 22 cardiologists and 21 medical students assessed the probability of CHF for 40 case vignettes. Since the diagnoses were based on thorough investigations and cardiologists' judgements ("gold standard"), diagnostic accomplishment could be analysed. The variation was large regarding strategies and diagnostic accomplishment between individuals, but not between the groups. The most important cues were cardiac volume and pulmonary congestion. Using cluster analysis, three main strategies were identified. Cardiac volume dominated in the first, pulmonary congestion in the second, and in the third the weights were more evenly distributed. The first cluster, comprising a third of the participants, had the best diagnostic accomplishment. In Study III, the same data were analysed for characteristics of the case vignettes causing the most and the least diagnostic agreement among the participants. Increased cardiac volume and presence of atrial fibrillation contributed to the diagnostic agreement between the participants, as well as a larger number of cues indicative of CHF. The starting point for Study IV was the recommendation in the CJA literature to use representative case vignettes. The concept of representativity and its consequences for the construction of case vignettes were discussed. Two factors above all turned out to be problematic: the incomplete information in the patient records and the necessity of keeping the number of case vignettes low. These two factors necessitated compromises regarding, for example, the choice of cues and the number of cues. In Study V, 15 GPs judged six case vignettes, selected from Study II, and the data were analysed regarding how different kinds of information were used in the diagnostic judgements. Although echocardiography (not included in the previous studies) was the most frequently used information, it was not used in a third of the judgement situations. Cardiac volume and pulmonary congestion were also important information. Information about other relevant diseases was frequently used in the diagnostic reasoning, but this is not reflected in the guidelines.

Both of the two methodological approaches to diagnostic judgements and reasoning in this thesis are useful tools for studying clinical decision-making. One possible application area is the study of expert doctors and medical students, which can give insights useful for teaching. Other application areas involve the development and testing of different decision support systems integrated in electronic patient records, and the development of guidelines.

Key words: case vignettes, Clinical Judgement Analysis, decision-making, general practice, guidelines, heart failure, judgements, think-aloud.

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to by their Roman numerals.

I. Skånér Y, Strender L-E, Bring J. How do GPs use clinical information in their judgements of heart failure? A Clinical Judgement Analysis study.
Scand J Prim Health Care 1998;16:95-100.

II. Skånér Y, Bring J, Ullman B, Strender L-E. The use of clinical information in diagnosing chronic heart failure: A comparison between general practitioners, cardiologists, and students.
J Clin Epidemiol 2000;53:1081-1088.

III. Skånér Y, Bring J, Ullman B, Strender L-E. Heart failure diagnosis in primary health care: clinical characteristics of problematic patients. A clinical judgement analysis study.
BMC Family Practice 2003;4:12.

IV. Skånér Y, Bring J, Strender L-E. Selecting representative case vignettes for clinical judgement studies: Examples from two heart failure studies.
Quality & Quantity, International Journal of Methodology (accepted for publication).

V. Skånér Y, Backlund L, Montgomery H, Bring J, Strender L-E. General practitioners' reasoning when considering the diagnosis heart failure: A comparison with guidelines.
Submitted.

CONTENTS

LIST OF ABBREVIATIONS

ACE	Angiotensin converting enzyme
CHF	Chronic heart failure
CJA	Clinical judgement analysis
CME	Continuing medical education
CV	Case vignette
ECG	Electrocardiography
ECHO	Echocardiography
EF / LVEF	Ejection fraction / Left ventricular ejection fraction
GP	General practitioner
IQR	Interquartile range
NPV	Negative predictive value
PHC	Primary health care
PPV	Positive predictive value
ROC	Receiver-operating characteristics

ERRATA IN THE ARTICLES

Paper	Page		Text:	Should be read:
I	98	Left column, row 16:	Table II	Table III
I	99	Right column, row 5:	Table II	Table III
II	1084	Left column, last row:	$\hat{y}^s$	$\hat{y}_s$
II	1085	Right column, row 6:	ecological validity	content validity
II	1085	Table 2; column: difference %-units, row: cluster strategy 2:	17	18
III	2	Methods, 2nd §, row 1:	physicians	participants

INTRODUCTION

THE FIELD OF STUDY

Decision-making is an essential part of health care. Physicians determine what is wrong with the patients and chose and recommend treatments. Patients decide whether to seek medical care and whether to follow treatment recommendations. Health policy makers decide upon the planning and financing of health care. All of these decisions determine the quality of the health care that is provided (Chapman & Sonnenberg, 2000). This thesis is about physicians' decisions and judgements.

Heart failure is common, it has a high morbidity and mortality, its prevalence is increasing, and it has a major impact on the health care system in Sweden and other European countries due to rising care costs (McMurray & Stewart, 2000; Mejhert et al., 2001). Heart failure is the most common reason for hospitalisation in patients over 65 years of age. The total health care cost for the management of heart failure in Sweden has been calculated at approximately SEK 2500 million (1996), which amounts to 2% of the total national health care budget (Bergsten-Rydén & Andersson, 1999).

In accordance with national guidelines, the majority of heart failure patients are managed by primary care physicians, i.e. general practitioners (GPs). The average heart failure patient makes three to four outpatient visits annually to his or her GP, and heart failure is one of the 10 most common reasons for visits to GPs (Cline et al, 2002; Malmberg & Persson, 2000). Of patients hospitalised for heart failure, 65% receive long-term follow-up in primary health care (PHC) (Persson et al., 1994). GPs thus play an important role in the management of the heart failure patient as regards diagnostics, treatment, and interaction with other caregivers. The focus of this thesis is on the GPs' diagnostic judgements of patients with suspected heart failure.

The judgements were made in response to written patient descriptions, or case vignettes (CV), that were based on authentic patients from PHC (Studies I, IV) or patients referred by GPs to a cardiology outpatient clinic (Studies II, III, IV, V). A decision is typically a choice between two or more alternatives. A judgement is a reasoning and deliberating process that may or may not lead to a decision. However, the distinction between these two concepts is not always clear, and in this thesis the two terms will be used in an overlapping way when describing the diagnostic process.

The complexity of the clinical work can be described as follows: "Uncertainty creeps into medical practice through every pore. Whether a physician is defining a disease, making a diagnosis, selecting a procedure, observing outcomes, assessing probabilities, assigning preferences, or putting it all together, he is walking on very slippery terrain. It is difficult for nonphysicians, and for many physicians, to appreciate how complex these tasks are, how poorly we understand them, and how easy it is for honest people to come to different conclusions" (Eddy, 1984).

HEART FAILURE

What is heart failure?

The term 'heart failure' is found in textbooks, classification lists, and scientific articles, and we can find information about the prevalence, recommended treatments and prognosis of heart failure. Nevertheless, it is an evasive concept. Not a distinct disease, it is rather an end stage of all diseases of the heart. It is often looked upon as a syndrome, or even a cluster of syndromes (Cleland & Habib, 1996). A modern description of heart failure which captures the complexity of the condition is: "Heart failure is a multisystem disorder which is characterised by abnormalities of cardiac, skeletal muscle, and renal function; stimulation of the sympathetic nervous system; and a complex pattern of neurohormonal changes." (Jackson et al., 2000)

There is no unequivocal, generally accepted definition of heart failure, but rather an abundance of definitions (Davis et al, 2000; Denolin et al., 1983; Marantz et al., 1988; Mosterd et al., 1997b; Wilhelmsen et al., 1989). These may include references to causal criteria (some kind of heart disease), to functional criteria (low cardiac output), to clinical criteria (characteristic symptoms and signs), to prognostic criteria (high mortality), and/or to therapeutic criteria (treatment targeting heart failure that results in clinical improvement).

In different contexts, and for different purposes, various definitions have been used, and this is one reason for differences regarding prevalence rates. In a PHC study in England on patients considered to have heart failure (on the basis of treatment with diuretics and some additional criteria) who underwent further examination for heart failure, the prevalence rate for those under 65 of age was found to be 0.6 cases/1000 and for those over 65 it was 28 cases/1000 (Parameshwar et al., 1992). In a Swedish, population-based study based on dyspnoea, with clinical criteria indicating heart failure (classified as latent or manifest), the prevalence rate of manifest congestive heart failure was 13% of the male population aged 67 (Eriksson et al., 1988). Recent studies of the epidemiology of heart failure in the United Kingdom are reported to use different criteria, for example different reference values for ejection fraction, for inclusion in the studies (Davis et al., 2000).

Heart failure can be specified as right or left, acute or chronic, systolic or diastolic. Right and left heart failure refer to syndromes presenting predominantly with congestion of the systemic or pulmonary veins, respectively. The terms do not necessarily indicate which ventricle is most severely damaged (The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology, 2001). Acute heart failure refers to a condition of acute cardiogenic dyspnoea characterised by signs of pulmonary congestion including pulmonary oedema, or cardiogenic shock characterised by a low arterial pressure, oliguria and a cool periphery (The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology, 2001). Chronic heart failure (CHF) is the most common form of heart failure. This thesis deals with CHF.

As ischaemic heart disease is the most common cause of heart failure in industrialised societies, most heart failure is associated with evidence of left ventricular systolic dysfunction. Left ventricular systolic dysfunction, as measured by the left ventricular ejection fraction (LVEF or EF) on echocardiography (ECHO), is often considered a gold standard for diagnosing heart failure, i.e. systolic heart failure, and modern heart failure treatment has been based on studies of this group of patients. However, there is a group of patients who have clinical findings corresponding to heart failure, but preserved left ventricular systolic function (normal EF), and they are presumed to have diastolic heart failure. This has undergone extensive debate in recent years, and there is growing interest in the diagnostics and treatment of diastolic heart failure (Caruana et al., 2000; European Study Group on Diastolic Heart Failure, 1998; Spencer & Lang, 1997; Zile, 2003; Zile & Brutsaert, 2002). The following definition has been suggested: "Diagnosis of primary diastolic heart failure requires three obligatory conditions to be simultaneously satisfied: 1) presence of signs and symptoms of CHF; 2) presence of normal or only mildly abnormal left ventricular systolic function; 3) evidence of abnormal left ventricular relaxation, filling, diastolic distensibility, or diastolic stiffness" (European Study Group on Diastolic Heart Failure, 1998). In a study by Zing et al, the conclusion was that the diagnosis of diastolic heart failure can be made without measurement of diastolic function if two criteria are present: 1) symptoms and signs of heart failure (Framingham criteria, see below), and 2) LVEF > 50% (Zile & Brutsaert, 2002).

Diagnosing heart failure

Prototypical CHF patients, as described in textbooks, often represent later stages of the disease, when there is already severe cardiac dysfunction. However, early detection of heart failure has become increasingly important, as modern drug treatment has the potential to improve symptoms and quality of life, slow down the rate of disease progression, and improve survival. Unfortunately, the clinical diagnosis of chronic heart failure is known to be difficult, especially in mild cases, as many features of the condition are not organ specific, and there may be few clinical features in the early stages of the disease. Most of the typical symptoms and signs are not specific for heart failure. One reason that CHF is difficult to diagnose is that it is not definable by an absolute level of any one parameter (Struthers, 2000).

Relatively few studies on patients suspected of having heart failure have been performed in primary health care settings. Those that have been conducted often report over-diagnosis (Clarke et al., 1994; Mair et al., 1996; Nielsen et al., 2001; Owens & Nease, 1997; Remes et al., 1991; Wheeldon et al., 1993). Some studies have indicated that as many as 30 – 70% of patients diagnosed in PHC as having CHF did not have the disease when investigated with objective investigations, mostly ECHO (Cowie et al., 1997; Owen & Cox, 2001; Remes et al., 1991; Wheeldon et al., 1993). However, some studies also report under-diagnosis (Blankfield et al., 1998; Hobbs et al., 2000; Morgan et al., 1999; Nielsen et al., 2001).

Symptoms and signs are important because they alert clinicians to the possibility of CHF as a diagnosis. However, they are not sufficiently specific for confirming left

ventricular systolic dysfunction (Davidson, 1996; Khunti et al., 2002; Timmis, 1996). Dyspnoea, ankle swelling and fatigue are often mentioned as the most typical symptoms. However, they are often difficult to interpret, particularly in obese patients and in women (Remes et al., 1991), and the inter-observer agreement between doctors may be low (Gadsboll et al., 1989). Peripheral oedema is a sign of CHF, but it has low predictive value, and it is often absent if treated or if the CHF is primarily a left ventricular dysfunction (Stevenson & Perloff, 1989). Neck vein distension is considered a good indicator of CHF, but the inter-observer agreement between non-specialists is fairly low, and it is not always found even in severe cases of CHF (Butman et al., 1993; Stevenson & Perloff, 1989). A third heart sound (gallop) is often considered specific for CHF, but can also have other causes, and the inter-observer agreement is often low (Folland et al., 1992; Ishmail et al., 1987; Stevenson & Perloff, 1989). The same can be said for pulmonary crepitations or rales (Spiteri et al., 1988).

The negative predictive value (NPV) of a normal electrocardiogram (ECG) in excluding left ventricular systolic dysfunction is more than 90%, and the use of the ECG as a screening investigation when it is difficult to get access to ECHO has also been proposed (Badgett et al., 1997; Mosterd et al., 1997a). However, patients with a normal ECG may also have left ventricular systolic dysfunction (Davie et al., 1996; Houghton et al., 1997). Cardiac enlargement and pulmonary congestion on chest X-ray are useful indicators of CHF, but not suitable as the only basis for therapy (Badgett et al., 1996). ECHO is recommended for getting objective evidence of cardiac dysfunction. The EF is often used as the gold standard for diagnosing systolic heart failure, although there is no definite value considered to be diagnostic. Different EF values have been used in clinical trials (Marantz et al., 1988). ECHO also gives information about possible causes of CHF, such as valvular diseases, ventricular hypertrophy, restricted ventricular motility, etc.

Natriuretic peptides have been suggested as a useful test for CHF in PHC (Smith et al., 2001). According to one study the positive predictive value (PPV) for symptomatic patients in PHC was 70%, while for community based screening it was only 16%, and the NPV in both cases was 98% (Struthers, 2000). It is recommended as a routine investigation for diagnosing CHF in the European guidelines (The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology, 2001), but it has not yet been incorporated as a routine investigation in Swedish practice.

Patients with CHF are often elderly, and co-morbidity is common. In a study in PHC in Stockholm, 98.7% of patients identified as having CHF had a co-morbidity. The most frequent were ischaemic heart disease (37.2%), hypertension (27.3%), chronic atrial fibrillation (23.7%), and diabetes (22.3%) (Nilsson & Strender, 2002). In a study in PHC in the UK, the combination of a history of myocardial infarction and a displaced apex (as a measurement of cardiac enlargement) had the best PPV for CHF (Davie et al., 1997).

Several scoring systems based on combinations of symptoms, signs and investigations (mostly chest X-ray) have been developed to assess the presence of CHF,

such as the Boston score, the Framingham score, and the men born in 1913 score (Carlson et al., 1985; McKee et al., 1971; Wilhelmsen et al., 1989). Six different scoring systems were investigated in a study (Mosterd et al., 1997). All scores had a high sensitivity for detecting definite CHF. The men born in 1913 (based on assessment of dyspnoea) gave a relatively large number of false positives. Although the scores were useful in detecting manifest CHF, objective measurements of cardiac function appear necessary to reduce the rate of false positives and to detect early stages of the disease.

Diagnostic guidelines

Guidelines have been defined as "systematically developed statements to assist practitioners and patient decisions about appropriate health care for specific clinical circumstances" (Field & Lohr, 1990). Improvement in quality of care and reduction of practice variation are common motives for introducing guidelines. Reducing the cost of health care can be a motive for guidelines (Aucott et al., 1994). The effect of guidelines on practice varies (Grimshaw & Russell, 1993). Attitudes change more often than practice (Lomas et al., 1989; Watkins et al., 1999). Obstacles to effectiveness may be lack of awareness of the guidelines, non-agreement with the content, self-efficacy, and difficulties in overcoming the inertia of previous practice (Cabana et al., 1999). Guidelines can be made more effective through incorporating doctors who are the targets of the guidelines in the development of the guidelines, through workshops and other interactive methods, and through continuing medical education (CME) (Borduas et al., 1998; Davis et al., 1995).

European guidelines for the diagnosis of heart failure were published 1995 by the European Society of Cardiology (The Task Force on Heart Failure of the European Society of Cardiology, 1995). The aim of the report was to provide practical guidelines for the diagnosis and assessment of heart failure for use in clinical practice, for epidemiological surveys and for clinical trials. Guidelines for the treatment of heart failure were published 1997 (The Task Force of the Working Group on Heart Failure of the European Society of Cardiology, 1997), and updated guidelines including both diagnosis and treatment in 2001 (The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology, 2001). A comprehensive version of the guidelines was published in 2002 (Remme et al., 2002). In Sweden, the Medical Products Agency published Swedish guidelines in 1996 based on the European version from 1995 (Medical Products Agency, 1996). Regional and local boards and committees have used the national guidelines in their work with locally adapted guidelines.

In Study V, we wanted to compare the GPs' diagnostic reasoning with the guidelines. We used the guidelines as a reference for evaluating the GPs' reasoning, but we did not test the influence of the guidelines on the diagnostic process. The diagnostic part of the guidelines consists mainly of reasoning text covering epidemiology, aetiology, pathophysiology, and possible methods for the diagnosis of heart failure in clinical practice, and it is difficult to use this text for assessment of diagnostic behaviour (Patel et al., 2001). Parts of the recommendations are more structured,

however, and they could be used as a reference, namely 1) a definition, listing some necessary conditions for the diagnosis CHF (Table 1), 2) a table, listing assessments to be performed routinely to establish the presence and likely causes of heart failure (Table 2), and 3) an algorithm for the diagnosis of CHF (Figure 1 in Study V).

Table 1. Definition of CHF from the European guidelines (2001). Criteria 1 and 2 should be fulfilled in all cases (The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology, 2001).

1.	Symptoms of heart failure (at rest or during exercise)
	and
2.	Objective evidence of cardiac dysfunction (at rest)
	and
3.	Response to treatment directed towards heart failure (in cases where the diagnosis is in doubt)

Table 2. Assessments to be performed routinely to establish the presence and likely cause of heart failure. From the European guidelines (2001). (Remme et al., 2002).

Assessments	The diagnosis of heart failure			Suggests alternative or additional diagnosis
	Necessary	Supports	Opposes	
Appropriate symptoms	+++		+++ (if absent)	
Appropriate signs		+++	+ (if absent)	
Cardiac dysfunctioning on imaging (usually echo-cardiography)	+++		+++ (if absent)	
Response of symptoms or signs to therapy		+++	+++ (if absent)	
Electrocardiography			+++ (if normal)	
Chest X-ray		+ (if pulmonary congestion *or* cardiomegaly)	+ (if normal)	Pulmonary disease
Full blood count*				Anaemia / secondary polycythaemia
Biochemistry and urinalysis*				Renal or hepatic disease / diabetes
Plasma concentration of natriuretic peptides in untreated patients (where available)*		+ (if elevated)	+++ (if normal)	

+ = of some importance; +++ = of great importance

*) Recommended assessments which are *not* included in the corresponding table in the Swedish guidelines, published in the Medical Products Agency version of the guidelines from 1996 (Medical Products Agency, 1996).

DOCTORS' JUDGEMENTS AND DECISIONS

Decision theory, a background

Decision-making has been a growing field of research during the second half of the 20th and the beginning of the 21st century. It is by nature interdisciplinary, and mathematicians, statisticians, economists, sociologists, military officers, psychologists and philosophers are among those who are studying it. Decision-making is also of great importance in medicine. It can be analysed according to different theories and research traditions, each with its own criteria for what is meant by a good or a correct decision. In classical decision analysis the criterion is rationality, defined as adherence to logic and to the probability calculus. In clinical judgement analysis, the criterion is agreement with conditions in the external reality (the ecology) and how well we can adapt to them. According to the naturalistic decision-making tradition, there is no better criterion than the opinion of experts in the field.

The study of decision-making can be characterised and divided according to different aspects. An important aspect is whether the study of decisions is *normative* or *descriptive*. Normative (or prescriptive) studies deal with how decisions should be made, and decision analysis, decision trees, and expected utility belong to this area. Descriptive studies deal with how decisions are made in real life. Another aspect is whether *individual* decisions are in focus, or decisions made by two or more persons interacting with each other, as in *game theory*, or whether the focus is on group decisions, as in *social choice theory*. Still another aspect is whether we talk about decisions under *certainty*, decisions under *risk*, or decisions under *uncertainty*. In this thesis, descriptive analyses have been used, individual decisions are studied, and the decisions and judgements involve uncertainty.

Much research has focused on the discrepancy between real decisions and decisions according to some norm for rational decision-making, as for example Bayes' theorem. In a study in 1954, Mehl showed that statistical treatment of available data often leads to better results than medical experts' clinical decisions (Meehl, 1954). Herbert Simon coined the expression "bounded rationality", meaning that people have a limited capacity for rational thinking, and instead develop simplified strategies, so called heuristics, in order to better deal with complex decision situations. (Simon, 1982). During the 1970s and 1980s, Kahneman and Tversky studied insufficiencies in human thinking in a systematic way in their research on heuristics and biases (Kahneman et al., 1982).

Many researchers have criticised this approach, and there seem to be two camps regarding the view of human rationality (Jungerman, 1983). One group, the "pessimists", claim that limitations in our cognitive capacity lead to systematic biases, while the other group, the "optimists", claim that human judgements in complex situations are on the whole effective, and that the violations of rationality norms found in empirical studies instead depend on the misleading use of normative models or other wrongly applied research methods. Other arguments could be that making perfectly rational decisions may be too costly (e.g. too time consuming) to be realistic, or that

decisions are not always single decisive acts, but rather links in a successive series of decisions, allowing for corrections over time as needed ("tree-felling" versus "hedge-clipping") (Connolly, 1988).

Clinical judgement analysis (CJA)

Most medical judgements are characterised by uncertainty. The connection between symptoms, signs, results of investigations, and diagnoses is often uncertain and probabilistic. Many clinical findings are based on subjective interpretations, and the reproducibility is low, as is the case, for example, with heart sounds and palpations (Gadsboll et al., 1989). Much information about a patient can suffer from a rather high degree of uncertainty in spite of being expressed in exact numbers as, for example, with cardiac volumes and ejection fraction values. Treatments do not lead with certainty to the expected results. In an environment characterised by uncertainty, it can also be difficult for us to learn from our experience (Brehmer, 1980). A theoretical basis for the study of people's adaptation to an uncertain environment is Brunswik's probabilistic functionalism and the Social Judgement Theory (Brehmer & Joyce, 1988; Cooksey, 1996; Leary, 1987). Judgement analysis is the term used for various methods employing multiple regression equations to model human judgement. When applied to medical judgements, the approach has often been called Clinical Judgement Analysis (CJA) (Kirwan et al., 1990). One of Brunswik's main points was that a method for studying judgements should provide simultaneous understanding both of human judgement itself and of the judgement task and context (the environment or the ecology), and this is captured in the symmetric lens model (Figure 1) (Brunswik, 1955).

In a typical CJA experiment, a doctor is asked to judge a number of patient cases (case vignettes) within a specified medical domain. The information in the case vignettes is varied systematically with regard to the value of a number of variables (cues), e.g. the patient's age, the presence or absence of symptoms and signs, or the values of laboratory tests. The judgement is often given as a probability, either as a number or as an answer on a graded response scale (e.g. the probability that the patient has a specific diagnosis). Interpreted in terms of the lens model in figure1, the left side, the environment (or "ecology"), represents the real state of the patient regarding the external criterion we are studying, e.g. a diagnosis, and the statistical relations (weights) between the cue values and this external criterion. The right side represents the subjective side, the judgement and the relations (weights) that the doctor assigns to the different cues in relation to this judgement.

The most common statistical method used in CJA is linear multiple regression, with the judgement as the dependent variable and the cues as independent variables, resulting in regression weights for each variable, which can be taken as representing the relative importance of this variable (cue) in determining the judgement. If the judgements are given in a dichotomised format (CHF or not, to treat or not to treat) instead of a probabilistic format, a logistic regression model could be used.

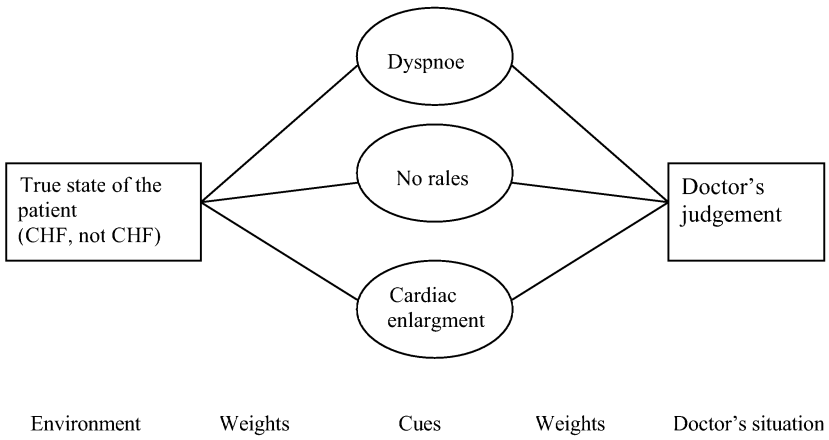


Figure 1. A simplified representation of the lens model applied to a medical judgement task.

The pattern of regression weights for each doctor in a CJA study is this doctor's strategy, often called judgement policy or decision policy. Policy capturing has been the focus of many studies in the medical field. It has been used for studying drug prescriptions (Backlund et al., 2000; Brown et al., 1997; Chaput de Saintonge & Hathaway, 1981; Elstein et al., 1986; González-Vallejo et al., 1998; Harries et al., 1996), diagnostic decisions (Braspenning & Sergeant, 1994; 1974; Tape et al., 1991; Vancheri et al., 2003; Wigton et al., 1986), and comparison between nations and between experts and novices (Kirwan et al., 1983c; Wigton, 1988). It has also been used as an educational tool, with the assumption that information about one's own policy in relation to the correct policy will lead to quicker learning (Klayman, 1988; Lundborg et al., 1999; Tape et al., 1992; Wigton et al., 1986). Using CJA for making policies explicit can also contribute to rational conflict solving (Chaput de Saintonge & Hattersley, 1985; Kirwan et al., 1983b). The participant's insight into his or her own judgement policies can also be assessed by comparing the regression weights and the participant's subjective opinion of the importance of the cues (Harries et al., 2000). Some of the general findings in CJA studies are that judges use few cues, they are often inconsistent, they usually have little insight regarding their own policies, and there are wide inter-individual differences between both judgements and policies (Brehmer & Brehmer, 1988).

In many studies only the right side of the lens model is investigated, often because the external criterion (the "true" state of the patient) on the left side is unknown. However, if there is a "gold standard", such as a reliable diagnosis, or a specific value representing the true state, judgements can be compared to the true state of the patient, and the doctors' policies can be compared with the ecological weights, which could then be interpreted as an "ideal" strategy (Cooksey, 1996). As an alternative, behaviour

according to guidelines or some other relevant norm can be chosen as an external criterion (Backlund et al., 2000).

Alternative models for analysing results from this kind of judgement task are the "fast and frugal" models like "Take The Best", based on the assumption that decisions are not necessarily based on linear compensatory integration of several sources of information, but instead are based on simpler rules and in some situations only on one cue (Dhami & Harries, 2001; Gigerenzer & Goldstein, 1996; Kee et al., 2003).

Representative case vignettes

The concept of representativity is fundamental to generalisation. Just as the subjects in an experiment must represent those not included in the experiment if generalisation over subjects is to be achieved, so also must the conditions of an experiment represent those conditions outside the laboratory over which generalisation is to be achieved (Brunswik, 1955, 1956; Wolf, 2000) In our studies, this means that not only the doctors, but also the sets of case vignettes, should be representative for the judgement situations we want to study. Not only should the distribution and range among environmental cues be represented, but also the intercorrelations. One reason for this is that to the degree that there is intercorrelation between cues, alternative strategies may work equally well, since cues may be substituted for one another. This is what is called "vicarious functioning" (Brehmer, 1994; Wolf, 1999). A representative design within the context of CJA has been defined in the following way: "The process of ensuring that naturally entangled and redundant aspects (…) of the ecology are not artificially disentangled for research purposes" (Cooksey, 1996).

An alternative to representative design is orthogonal design, which is a way of representing all cue combinations in a systematic way. This can be done by using all possible combinations of cues and cue levels for the construction of case vignettes, as for example in a study in which 12 vignettes were used, representing all the combinations of three cues, with three, two, and two levels, respectively (Elstein et al., 1986). Since the number of possible combinations can be very large, the number can also be reduced in a systematic way (Green, 1974). In a representative design, the cases are supposed to be representative of the environment, but in an orthogonal design they are supposed to be representative of the population of possible cue combinations. An advantage of the orthogonal design is that with the same number of case vignettes the statistical power will probably be better. However, for medical research, especially with experienced judges as participants, it will be advantageous to use a representative design where all the natural redundancies and intercorrelations are represented in the case vignettes (Cooksey, 1996). A further disadvantage of using artificially created cases is that there will be no external criterion that might be used as a gold standard for the judgements.

In Study IV we have described the requirements for creating sets of representative case vignettes, and used our experiences from Studies I and II to illustrate problems and possible consequences involved.

Selecting a relevant population of patients: The set of case vignettes should ideally be a sample of a specified population of patients corresponding to the patients the doctors normally meet, and reflect the characteristics of this population, in our case the frequency of CHF, the distribution and range of cue values, and the intercorrelations of the cues among their patients. However, representing the whole population or even the population of all patients (for example all patients listed at a health centre) would not be realistic, since too few relevant cases would be selected (Figure 2). An ideal population would rather be "patients with symptoms and signs that might indicate CHF". However, this will be difficult in practice, since patient records can seldom be searched for symptoms or signs.

Selecting relevant cues: According to Hammond, it is important to have a theory about which cues are important, i.e. the ones that "if ignored would produce a critically different result than if not ignored" (Hammond, 1998). This theory should include knowledge about the environment to an extent we seldom have, but also knowledge about the doctors' opinions of the value of the cues for the judgement, whether justified or not.

Deciding about the number of variables: From the perspective of the participants, a large amount of information would probably seem realistic. However, it would probably also be confusing. Since the number of cues will influence the number of case vignettes needed, this makes it important to keep the number as low as possible, without omitting any relevant cues.

Deciding about the number of case vignettes: The number of case vignettes needed is dependent on (*a*) the number of cues, (*b*) the intercorrelation between cues, (*c*) the variation of cue values in the set of case vignettes, (*d*) the intra-rater reliability of the participants (their consistency), (*e*) the differences to be detected, (*f*) the "endurance" of the participants. A large number of cues, a high intercorrelation between cues, a low variation in cue values and low intra-rater reliability will result in a low statistical power of the study, which can be compensated for by increasing the number of case vignettes. To detect small judgement differences among the participants, a larger number of case vignettes will be needed, and it is therefore useful to decide what difference would be of some clinical importance. Case vignettes in CJA studies tend to look rather repetitive, and too many vignettes may bore the participants, which may reduce the reliability of their judgements. A rule of thumb regarding the number of case vignettes is to have a ratio between the number of case vignettes and the number of cues that is between five to one and ten to one (Cooksey, 1996).

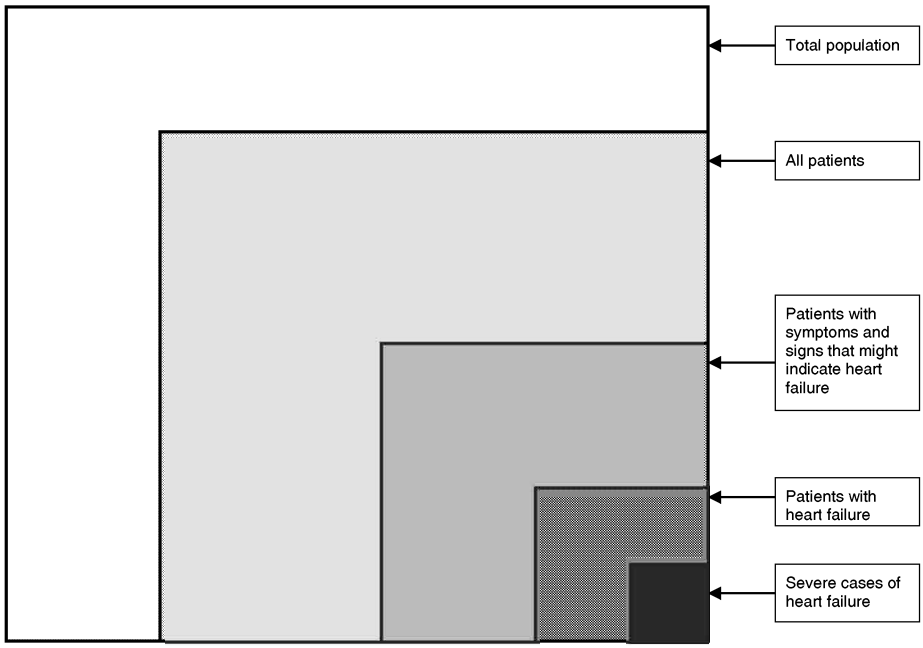


Figure 2. The figure shows the relationship between different populations: the total population, the population of patients, the population of patients with symptoms and signs that could indicate CHF, the population of patients with CHF, and the population of patients with a severe form of CHF.

The use of routine patients will reflect important aspects of the task, such as intercorrelations and redundancy among cues. However, the variation of cue values will probably be low (many "normal" cases), and together with high intercorrelations, this may result in low statistical power, which has to be compensated for by increasing the number of case vignettes. There may thus be a tension between representativity and statistical power. Methods for compensating low statistical power depending on low cue variation have been suggested, for example adding some cases with more "extreme" values (McClelland, 1999).

Other factors that can also be important, such as the mode of presentation, the procedure of selecting patient records, and the decision as to whether or not to inform the participants in a study about the sampling process, are treated in Study IV.

Structure and process aspects

In CJA studies, the regression weights are said to "capture" the policy or the judgement strategy. However, what is captured is the mathematical relation between input (cues) and outcome (judgements). This can be useful in many situations, for example when predicting judgements of new cases and for educational purposes. The CJA method has been said to describe the structure but not the process of the judgement, and for this

reason it has been called a "black box theory" (Elstein et al., 1978). Even if the linear regression models often fit data well and are looked upon as fairly robust models of judgement processes, they do not give us information about how information is processed, for example how and when hypotheses are generated, how judgements are changed over time in response to new information, and how medical knowledge is used.

In the process-tracing approach the focus is on the process itself, and not on the relation between input and outcomes of the process. Data are gathered during this process in order to reveal the train of thoughts leading to a final judgement. The process-tracing techniques are more related to problem-solving research, and the basic thought is that the cognitive processes, such as problem-solving or decision-making, should be studied by collecting data during the process (Svenson, 1979). One type of data that can be collected is data revealing the information search pattern. This has been done with various methods, for example by eye movement recording, by letting subjects turn cards with information, and more recently by computer logging. The assumptions here are that paying attention to an item reflects the processing of it at a cognitive level, and that a longer attention time can reflect a more complex cognitive process. Another type of data that has been commonly used for tracing processes is think-aloud protocols, which may produce rich verbal data about reasoning during a task (Fonteyn et al., 1993; Svenson, 1979).

The value of the data obtained from the think-aloud method is based on certain assumptions about verbalised data: a) the cognitive processes that generate verbalisations are a subset of the cognitive processes that generate any type of recordable response or behaviour, b) human cognition is information processing, a sequence of internal states successively transformed by a series of information processes, and c) information recently acquired and currently being concentrated on is directly accessible as verbal data, using the think-aloud method (Ericsson & Simon, 1984).

There has also been criticism of the think-aloud method, claiming that verbal reports might produce inaccurate data. One reason for that would be that many cognitive processes are unconscious, and therefore not possible to report about (Nisbett & Wilson, 1977; Russo et al., 1989). However, in a study of decision-making using the think-aloud method, no apparent effect of verbalisation was found when comparing three groups: think-aloud during the task, retrospective recall, and no verbalisation during the task (Henry et al., 1989). Other studies have helped to clarify which specific conditions might affect the validity of the verbal reports (Ericsson & Simon, 1999). In general, the concurrent use of think-aloud and performance of a cognitive task generates better verbal reporting than retrospective think-aloud. The instructions to the subjects are also important. Instructions simply to report their thoughts does not seem to distort their cognitive processes, while instructions requiring them to explain or justify their judgements may do so. Avoiding eye contact between researcher and subject is also important in order to control for non-verbal communication.

Think-aloud technique applied to medical decision-making

In the 1970s Elstein and co-workers studied medical diagnosing with the use of think-aloud technique, and their theoretical framework was the diagnosis as a hypothetical-deductive process (Elstein et al., 1978). One of their important conclusions, which has been much quoted, was that doctors tended to make tentative diagnostic hypotheses very early in the patient encounter. Another important conclusion concerned decision-making by subjects with different levels of expertise. They found that there were no differences between experts and novices regarding general problem-solving skills. The differences in accomplishment were instead related to differences in experts' and novices' representation of the knowledge in the domain. Experts generally have a more highly organised knowledge base, which allows them to partition the problem into manageable 'chunks' (Elstein et al., 1978; Patel et al., 2001).

Research about knowledge organisation and knowledge structure is closely related to the concept of 'scripts' or 'illness scripts' (Charlin et al., 2000; Custers et al., 1996, 1998; Lamond et al., 1996; Schmidt et al., 1990). A script can be described as the mental organisation of a disease or a medical procedure or some other limited medical area (Schmidt & Boshuizen, 1993). The generation of a diagnosis can then be described as the activation of a script. In most routine diagnostic situations, biomedical reasoning is not very important, and other factors such as the so-called enabling conditions characterise the scripts. Enabling conditions are background factors like age, sex, and previous diseases, which increase the likelihood of a certain disease.

Expert decision-making has been a major interest in decision research. One of the reasons for this is the expectation that expert knowledge could be 'extracted' from experts and put to use in computerized decision support systems, for example. Expertise research has shown that there are fundamental differences in comprehension, problem-solving, and decision-making as a function of expertise (Chi et al., 1988; Ericsson & Smith, 1991; Patel et al., 2001). Some of the differences between experts and novices that have been described, in addition to their above-mentioned different knowledge organisation, are the following: 1) Experts and non-experts use different patterns of reasoning. In routine problems, experts use a data-driven pattern of reasoning, while non-experts use a hypothesis-driven pattern of reasoning where the hypothesis guides data collection and interpretation. 2) Experts tend to make errors of omission because of over-confidence, while non-experts tend to make errors of commission, because of their lesser competence in discriminating relevant from irrelevant information. 3) Experts tend to generate a small set of relevant hypotheses at a high level of abstraction that can quickly be narrowed down, and tend to use rules of thumb for finding an approximately correct solution relatively fast, while non-experts must take a longer time for their decisions. 4) Experts interpret information about cases more according to their experience than according to scientific evidence.

Some of these findings may have implications for education. One study shows that the better diagnostic accomplishment reached by experts was related to their use of pattern recognition and illness scripts, and recommends that this be taught to students (Coderre et al., 2003), and another study looked for expert heuristics, which could also

be used in the education of students (Fisher & Fonteyn, 1995). GPs' use of simple heuristics and rules of thumb has been investigated in several studies (Andre et al., 2002; Andre et al., 2003; Essex & Healy, 1994).

This knowledge also has relevance for the construction of guidelines and computerised decision support systems. These are often written by experts for non-experts, and if their knowledge structures differ, this may cause problems and make the use of guidelines and decision support problematic (Kushniruk et al., 1995).

AIMS

GENERAL AIM

The general aim of the studies was to increase our understanding of how Swedish GPs make diagnostic judgements concerning patients with suspected CHF. Increased knowledge in this area will enhance our possibilities to improve guidelines and teaching regarding CHF diagnostics, and to design training and education. It will also be useful for developing decision support in this area.

SPECIFIC AIMS

Study I
To describe GPs' diagnostic strategies in terms of which factors influence their diagnostic judgements of patients with suspected CHF. To describe the variation among GPs regarding diagnostic strategies and diagnostic judgements.

Study II
To describe and compare GPs', cardiologists', and medical students' diagnostic strategies and diagnostic judgements when assessing patients with suspected CHF. To compare the participants' strategies with the optimal diagnostic strategy, given the available information. To compare the participants' diagnostic accomplishment.

Study III
To analyse how patient characteristics can contribute to difficulties in diagnosing CHF. Our hypothesis was that typical CHF cases (many cues indicating CHF) and typical non-CHF patients (few or no variables indicating CHF) would be easier for the participants to agree about than the intermediate cases.

Study IV
To discuss the concept of representativity and problems concerning the construction of sets of representative case vignettes for use in CJA studies, with examples taken from our studies.

Study V
To describe, by using think-aloud data, how GPs' diagnostic reasoning and diagnostic judgements about patients with suspected CHF are related to the recommendations in the European and Swedish guidelines.

METHODS

CASE VIGNETTES

Written case vignettes were used in studies I, II, III, and V. In Study IV, theoretical and practical problems regarding representative case vignettes were discussed.

The cues to be presented in the CJA studies (I-III) were chosen because of their relevance for diagnosing CHF according to articles, textbooks, interviews with GPs, and prediction validity in relevant populations, and because of their availability in the patients' medical records.

Study I

The participants judged 45 case vignettes, five of which were duplicates. The case vignettes were based on authentic patients, selected from two health centres in Stockholm. All the patients had the diagnosis CHF. One reason for this was that otherwise it turned out to be difficult to find information about the variables of interest in the medical records. However, the variation in the distribution of variable values in the set of case vignettes was large enough to represent both patients with low and with high probability of CHF. It was also large enough to make it possible to estimate the regression coefficients with the desired statistical precision. Intercorrelations between the variables were low.

Information about ten cues was presented in the vignettes: age, sex, history of myocardial infarction, dyspnoea, peripheral oedema, lung auscultation, cardiac rhythm, heart rate, cardiac volume and pulmonary congestion. Information about ECHO was not included, since we wanted to capture an earlier step in the judgement process. The cues and the cue values (levels) are listed in Table 3.

Study II

The participants judged 40 case vignettes, based on patients referred by GPs to the cardiology out-patient clinic at the Södra Hospital in Stockholm during the period 1993-1995 for problems related to heart failure. One reason for selecting cases from the cardiology department was that we wanted to use the diagnoses as gold standards for the participants' diagnostic judgements, and we considered the cardiologists' diagnoses, based on all available information, to be the best we could get. We also wanted to include both cases with and without CHF. Here, too, the possibility of finding information about the relevant variables in the medical records was a limiting factor.

Table 3. Clinical information given in the case vignettes in Study I.

Cue	Cue value
1. Age	60-70 years
	80-90 years
2. Sex	Male
	Female
3. History of myocardial infarction (MI)	No MI
	Had an MI more than a year ago
4. Dyspnoea	No dyspnoea
	Dyspnoea when climbing
	Dyspnoea when walking on level ground
5. Oedema	No leg oedema
	Moderate leg oedema
6. Lung auscultation	No rales, no rhonchi
	Basal rales, no rhonchi
7. Cardiac rhythm	Regular rhythm
	Irregular rhythm
8. Heart rate	≤ 90 beats/min
	≥ 110 beats/min
9. Heart X-ray	No cardiac enlargement
	Cardiac enlargement
10. Lung X-ray	No signs of stasis
	Stasis

Information about eight cues was presented in the vignettes: history of myocardial infarction, atrial fibrillation, dyspnoea, peripheral oedema, rales, systolic blood pressure, cardiac volume and pulmonary congestion. Age and sex were presented in the case vignettes, without being included in the model, since they had turned out to be of very little significance for the judgements in Study I (within the age range in the study). A low blood pressure was found to be associated with the diagnosis CHF among the group of patients used for constructing the case vignettes, and information about systolic blood pressure was therefore added. ECHO was not included, for the same reason as in Study I. Twenty-six of the cases had CHF, and 14 did not. Figure 3 shows an example of a case vignette.

The cue levels were presented according to the nature of the cue, in order to be as realistic as possible, within the limits of the method. Age, blood pressure, and relative cardiac volume were thus presented as number of years, mm Hg, ml/m^2, respectively. The rest of the cues were categorical with two levels (dyspnoea on three levels).

Study III

The same cases were used as in Study II. We specifically analysed those case vignettes that in Study II had been the ten most problematic cases (the largest diagnostic disagreement among the participants) and the ten least problematic cases (the smallest diagnostic disagreement among the participants).

The patient is a 72-year-old woman who consults you for increasing fatigue. She suffers from breathlessness when walking on level ground. She has no history of myocardial infarction. She has atrial fibrillation. No dependent oedema. She is being treated with diuretics.

On examination
Heart Arrhythmia, no tachycardia. Systolic blood pressure 110 mm Hg
Lungs Basal rales, no rhonchi

You also get to know
Heart X-ray: Relative heart volume 820 ml/m^2
Lung X-ray: Minor signs of pulmonary stasis
Hb, white blood cell count, electrolytes and TSH are normal

Your judgement of the patient
What is your assessment of the probability that this patient has some degree of heart failure?
(Put a cross on the line corresponding to the probability that corresponds to your assessment.)

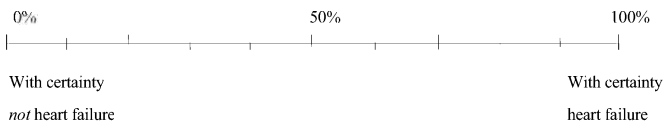


Figure 3. Example of a case vignette from Study II.

Study IV

The case vignettes from Studies I and II were used to illustrate problems that can arise when selecting representative cases for a CJA study.

Study V

Six cases (CV1 – CV6) were selected strategically from the cases used in Study II. They were selected to represent patients with and without CHF, and with various degrees of difficulty. In this study the case vignettes contained more information about the patients, for example about other relevant diseases (e.g. coronary heart disease, stroke, diabetes), lifestyle factors (e.g. smoking, alcohol consumption), symptoms,

signs, ECG, and chest X-ray findings. Because the information was presented successively to the participants in this study, ECHO results were included. Fifty variables were presented in the set of six case vignettes, 19 of them were included in all six vignettes, 17 only in one. For each case vignette, the presented variables were listed, and coded for content and value (Table 4).

Table 4. Examples of coding of variables in the case vignettes. Positive value = presence of finding or pathological finding, negative value = absence of finding or normal finding.

Information as presented in the case vignette	Content	Value
"Shortness of breath when walking on level ground"	dyspnoea	Positive (presence of finding)
"Pathological R-progression on ECG"	ECG	Positive (pathology)
"He has not had swollen legs"	oedema	Negative (absence of finding)
"Regular rhythm"	rhythm	Negative (normality)
"Relative heart volume 630 ml/m^2"	heart volume	630 (numeric values as presented in the text)

PARTICIPANTS

Study I

We wanted 25-30 GPs, specialists in family practice, to participate. Eleven Continuous Medical Education (CME) groups in the northern part of Stockholm were invited to take part in the study. In order to supplement the number of GPs we could recruit from the CMEs, we also invited GPs from 13 health centres in the same area. The GPs were not randomly selected.

Studies II and III

We wanted GPs, cardiologists, and medical students to participate, about 25 in each group. Fifty GPs, selected randomly from a list of specialists in family medicine in the southern part of Stockholm (n = 332), were invited to participate in the study. Only GPs who were specialists in family medicine were included. All cardiologists (n = 38) from two cardiology clinics in Stockholm, and all medical students from two courses in family medicine (n = 82), were also invited to participate. Studies II and III were based on the same original data.

Study IV

The same participants as in Studies I and II were used in this study.

Study V

We wanted 15 GPs to participate. All health care centres in northern Stockholm within a distance of 20-30 kilometres from the city centre (n = 61) were listed and contacted

in a random order. In each health centre the GPs were contacted in random order. Only one GP at each health centre was included in the study, and this person had to be a specialist in family medicine.

DESIGN, PROCEDURE AND DATA ANALYSES

Study I

This study used the CJA approach to compare the relative weights given to different kinds of information (cues) about the patient in the judgements of the probability of CHF. The 45 case vignettes were presented to the GPs at practice visits. For each vignette, the GPs were asked to assess the probability that the patient suffered from any degree of CHF. The assessments were made on a visual analogue scale with "totally unlikely" at one end (0%), and "certain" (100%) at the other. The response scale was similar to the one in Figure 3, except that the line was not divided in percentiles.

After the GPs had completed the case vignettes, they filled in a questionnaire with some questions about themselves and about the task. They were also asked to mark and rank the four cues that they themselves thought had been most important for their judgements. All the answers were given anonymously, marked only with a code number, which was used for feedback purposes only; the codes were not used by us to identify the participating GPs.

Each individual GP's judgements were analysed separately. In a linear multiple regression model, the probability assessments were used as dependent variables, and the cue values as independent variables. The regression coefficients were used as a measurement of the importance of each cue for the diagnostic judgements. The magnitude of the regression coefficient for a cue shows the amount by which the probability of CHF increases in the opinion of a particular doctor when the cue is present/pathological (or in case of absence/normality, how much the probability decreases).

All the cues except dyspnoea were dichotomous, which makes the interpretation of the regression coefficients rather straightforward. If two case vignettes are identical regarding all cues except oedema, and the regression coefficient for oedema is 2.0, this means that having oedema increases the probability of CHF by 2%, compared with not having it. As for dyspnoea, we used the average effect of going from no dyspnoea to dyspnoea when climbing, and from dyspnoea when climbing to dyspnoea when walking on level ground.

For each vignette the median value of the participants' assessments was calculated. A high median value indicates that the participants consider it probable that the patient has CHF.

Study II

In this study we used the CJA approach to compare three groups of participants regarding judgements and strategies: GPs, cardiologists, and medical students. We also wanted to compare the participants' judgements with a gold standard diagnosis, and their judgemental strategies with an optimal strategy, given the information available in the case vignettes. Results from Study I were used for planning Study II regarding selection of cues (see above), number of participants, and calculation of power. Expecting the intra-rater reliability to be about as good as in Study I (R^2 = 70), and considering a regression coefficient difference of about 15% to be of clinical interest, the power was calculated to be 0.82. The good intra-rater reliability in Study I made us refrain from duplicate cases in order to increase the power of this study.

The case vignettes were presented in a booklet, which was sent to the GPs' work addresses, distributed to the cardiologists by one of the authors (BU), and distributed to the students at the beginning of their two-week course in family medicine. For each vignette, the participants were asked to assess the probability that the patient suffered from any degree of CHF, in the same way as in study I (Figure 3). For each vignette the median value of the participants' assessments was calculated. A high median value indicates that the participants consider it probable that the patient has CHF.

For each participant, we made a linear multiple regression model. The probability assessments were used as dependent variables, and the cue values as independent variables. Analysis of continuous variables using natural units would result in regression coefficients reflecting the influence on the judgements of a change of one unit (one year of age, one mm Hg, one ml/m^2). This would not be clinically relevant, and we therefore modified the scale steps to make the regression coefficients reflect the effect of a clinically meaningful change of the values: for age 20 years, for systolic blood pressure 30 mm Hg, and for relative heart volume 300 ml/m^2.

For the group comparisons, the aggregated strategies for each group were calculated as the means of the regression coefficients for the members of each respective group (Cooksey, 1996). Since mean-value calculation may disguise interesting subgroups within the three groups, we also used cluster analysis across all participants to look for specific patterns with regard to cue utilisation (Johnson & Wichern, 1988).

A reference model was constructed, with the clinical diagnosis as the dependent variable and the cue values as independent variables. This corresponds to the left side of the lens model, while the participants' judgements correspond to the right side (Figure 4). The reference model was supposed to represent a strategy that makes optimal use of the information presented in the case vignettes. It was used for deriving a prediction (event probability) for each case ($\hat{y}_e$), which was used for evaluating the participants' judgements. Alternative reference models based on subsets of the cues were also constructed, and specificity and sensitivity for those models were calculated. For each participant, the linear regression model was used to derive the predicted judgement for each patient ($\hat{y}_s$). For each of the three groups, the average predicted

judgement was calculated, and correlations between the average predicted judgements for the three groups, and between those and the event probability of the "optimal strategy", were calculated.

The diagnostic accomplishment could not be assessed by a direct comparison for each case, since the participants' judgements were made as probability assessments (0-100%), while the real patients had CHF or not. Ideally, all the cases with CHF should be considered to have a probability of CHF of 100%, and all the cases with non-CHF should be considered to have a probability of CHF of 0%. An effective judgement strategy should thus maximise the difference between the two groups of cases. The average event probability ($\hat{y}_e$) for cases with CHF was 84%, and for cases without CHF 29%, and the difference, 55 percentage units, was considered to be the best diagnostic accomplishment that could be expected by the participants, given the information presented in the case vignettes (Figure 4).

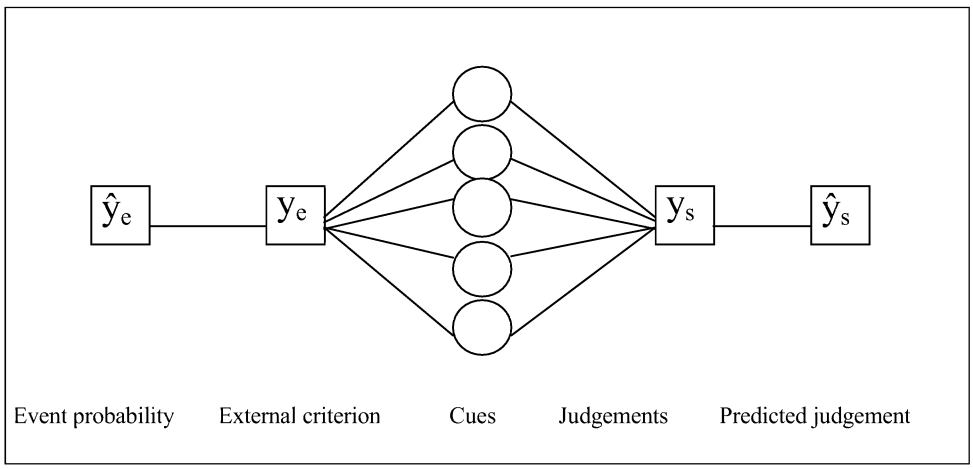


Figure 4. In this extended lens model, y_e represents the external criterion, the clinical diagnosis (CHF or non-CHF), and $\hat{y}_e$ represents the probability for CHF derived from the optimal model, i.e. the average prognosticated probability of CHF. On the right side of the model, y_s represents the participants' judgements, and $\hat{y}_s$ represents the judgements predicted by the participants' regression model.

Study III

Case vignettes, participants, procedures, and some of the analyses were the same as described in Study II, but in this study we focused on patient characteristics. In Study II we could not demonstrate any differences between the three groups of participants regarding judgements and judgement strategies, and therefore all participants could be treated as one group in the context of this study.

The ten most problematic cases (those giving rise to the most diverging probability assessments) and the ten least problematic cases in Study II (those giving rise to the

least diverging probability assessments) were identified and analysed separately. As a measurement of divergence, we used the interquartile range (IQR, the difference between the third and the first quartile) of the participants' assessments.

In order to count the number of cues indicative of CHF for each case vignette, we dichotomised the non-binary cues. A positive cue for systolic blood pressure was ≤ 140 mm Hg (since a low value had been found to be associated with heart failure in our group of patients), and for relative cardiac volume a positive cue was ≥ 490 ml/m^2 for men and ≥ 450 ml/m^2 for women (corresponding to the local reference values). The association between the interquartile ranges and the number of positive cues for all the vignettes was studied with a regression line plot.

Study IV

In the CJA tradition there are strong arguments in favour of using representative case vignettes. Creating a set of representative case vignettes meeting all theoretical requirements will often be difficult. Selecting a relevant population and relevant variables, deciding about the appropriate number of variables and case vignettes, and deciding whether case vignettes based on authentic patients or artificially constructed case vignettes should be used are some of the problems that must be dealt with. In this study, the two sets of case vignettes in Studies I and II are compared and discussed with reference to our aim of studying the CHF diagnostics in PHC with CJA: which compromises are necessary, and which consequences might follow from these compromises.

Study V

The GPs' use of clinical information as diagnostic arguments was compared with the recommendations in the guidelines. The GPs' diagnostic judgements were compared with the clinical diagnoses.

Before the study, the GPs had received written information about the aim of the study (to study clinical judgements) and about the method (think-aloud), but not about what kind of medical problems that would be presented to them. The study was conducted at the GPs' offices. The participants were instructed that six authentic patients, suspected by GPs to have CHF, would be presented, and that their task was to say aloud their thoughts about the cases, and try to decide whether the patients had CHF or not. The GPs received the same six case vignettes and in the same order, so as to reduce the variance due to differences in case order. The different kinds of information were presented in the same order for all the six cases. Each vignette was presented on a computer screen in five successive steps, and the doctors controlled the shift to a new screen by clicking with the mouse on a continue button. On a sixth screen, they were asked to summarise their judgements about the case and try to decide about the diagnosis. The GPs could express their diagnostic judgements freely, with their own words. Time spent on each screen was automatically logged. For the computer presentation, QA™ software was used (Montgomery & Hammarberg, 2000).

The researcher was seated behind the doctor during the session. A participant who was silent for more than about 15 seconds was reminded to say his or her thoughts aloud about the information presented (Ericsson & Simon, 1999). Before the session started, the participants got a "test case" (not recorded) in order to get acquainted with the method. The think-aloud sessions were tape-recorded and transcribed by a secretary.

The protocols were searched for mentionings of case variables, and each proposition (or segment of a proposition) containing a variable was coded for content and value; *content* meaning the variable that was used, and *value* meaning the direction of the argument, i.e. whether the GP seemed to use the variable as an argument *for* the diagnosis of CHF, as an argument *against* CHF, or as not being of any explicit use for the diagnosis ("*mentioning only*"). For each participant, a specific evaluation of each variable was only counted once for each case vignette, in order not to give more weight to thoughtful repetitions of an argument than to a single, firm statement. However, if a participant used the same variable as an argument both for and against the diagnosis CHF, both evaluations were coded. Ten percent of the 90 case vignette protocols were selected at random and coded independently by two of the authors (YS, LB) to estimate the inter-rater agreement of the coding.

The participants were not forced to express their diagnostic judgements in a specific format, and their free verbal statements therefore had to be interpreted and coded. Two of the authors (YS, LB) independently classified all the diagnostic judgements in three categories: CHF or probable CHF; uncertainty about the diagnosis; not CHF or probably not CHF.

To make it possible to compare the GPs' use of case vignette information as diagnostic arguments with the recommendations in the guidelines, we used the simplified recommendations, which were condensed in a definition (Table 1), and a table of assessments to be performed routinely to establish the presence of CHF (Table 2) (Remme et al., 2002).

RESULTS

STUDY I

Participants

The participants were not selected from a well-defined population of GPs, and a response rate can therefore not be obtained. Twenty-seven GPs participated in the study, 14 men and 13 women. They had worked as GPs on average 9 years (range 1.5-22 years).

The GPs' judgement strategies

On an aggregated level, the three highest regression coefficients were lung X-ray (33.5, range 1.3-70.9), heart X-ray (16.9, range –5.9-44.5), and history of myocardial infarction (9.5, range -14.4-19.7) (Figure 5). For 22 GPs, the highest regression coefficient was lung X-ray, and for five GPs it was heart X-ray.

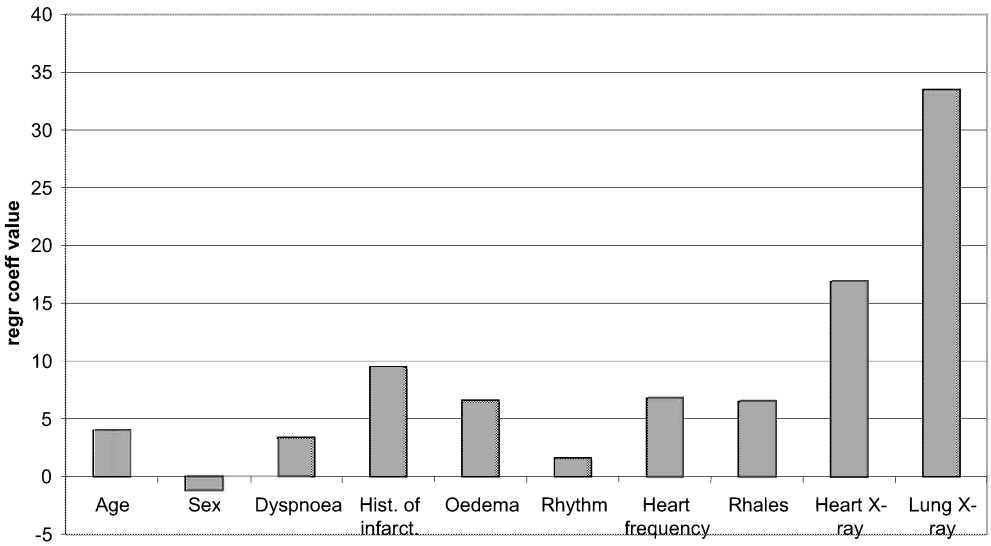


Figure 5. The mean regression coefficient values for the group of GPs. (The negative value for sex is an effect of men being coded as 1 and women as 2.)

According to the GPs' own stated opinions about their strategies, 20 GPs considered lung X-ray as being one of the three most important cues for their judgements, while it had among the three largest regression coefficients for 24 GPs. The corresponding numbers were 12 (19) for heart X-ray, 14 (2) for dyspnoea, and 8 (14) for history of myocardial infarction. The GPs thus had a fairly good idea of their own judgement strategies, except that they overestimated the importance of information about dyspnoea for their judgements.

The median adjusted coefficient of determination (R^2 adj) was 70 (range 30-96), which indicates that the regression model predicts the judgements well.

The GPs' assessments of the probability of heart failure

The distributions of probability assessments for the individual cases varied among the GPs. The mean values for the GPs for the whole set of case vignettes varied between 39 and 76 percentage units (standard deviations 15-41). In figure 6, three examples of GP assessment distributions are shown: GP6 tends to assess the probability of CHF as low for most of the cases, GP16 tends to assess it as high, while GP12 divides the cases in two groups, one with low and one with high probability of CHF.

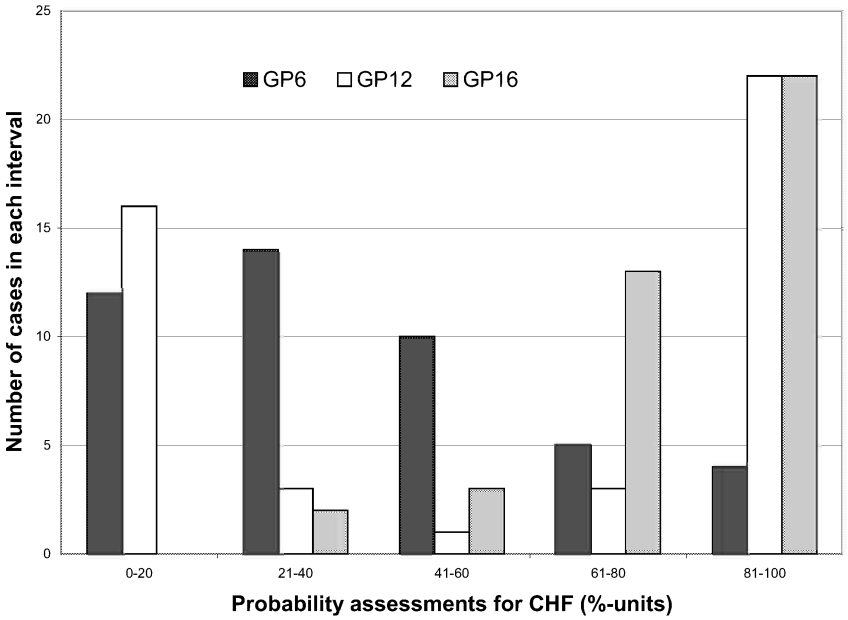


Figure 6. Three GPs' distributions of probability assessments.

There was also a large variation between the GPs' probability assessments of the individual case vignettes, i.e. a high inter-individual variation. The greatest range between different GPs' assessments of an individual case was 94, and the smallest was 19 percentage units. However, the intra-individual variation was small, as measured by the assessments of the duplicate cases. We made 134 observations (27 GPs judging 5 pairs of cases; missing data for one pair), and in 62% of them the difference between the assessments of the paired cases was as low as 0-10 percentage units, in 25% of them it was 11-20, and in 13% more than 20. The participants thus seemed to be consistent in their judgements.

The GPs' opinion of the representativity of the case vignettes

In the questionnaire, the GPs were asked if they thought the case vignettes could have been collected from their own practice. Twenty-four said yes, one said no, and for two GPs data were missing.

STUDY II

Participants, response rate

Twenty-seven GPs (14 men, 13 women) and 22 cardiologists (17 men, 5 women) participated in the study. The average age of the GPs was 48 years and for cardiologists it was 50 years. The average time as a GP was 10 years, and as a specialist in cardiology 11 years. We found no significant differences between participants and non-participants regarding age and sex for GPs and cardiologists.

Twenty-one medical students also participated in the study (7 men, 13 women, missing information about sex for one student). The average age of the students was 28 years. We found no significant differences between participating students and the whole group of students regarding age and sex.

The response rate was 54% for the GPs, 58% for the cardiologists, and 26% for the medical students.

Judgement strategies

The judgement strategies were similar for GPs, cardiologists and students (Figure 7). Cardiac enlargement was the most influential cue (the average regression coefficient was 21, 22, and 19 for GPs, cardiologists and students, respectively). Because of our coding system, this meant that if a patient had a heart volume that was 300 ml/m^2 larger than another patient, but otherwise had the same cue values, this patient would be considered to have a 21%, 22%, or 19% greater risk of having heart failure than the other patient. For the rest of the cue values for the three groups, see Figure 7. Sex and age did not influence the judgements.

By cluster analysis we identified three main strategies among the participants. In the first cluster (n = 22) cardiac volume was the dominant cue, in the second cluster (n = 12) pulmonary congestion was the dominant cue, and in the third cluster (n = 36) the weights were more evenly distributed.

The correlations between the average predicted judgements of the three groups, when compared in pairs, were high: between GPs and cardiologists 0.99, between GPs and students 0.95, and between cardiologists and students 0.95. The correlations between the average predicted judgements of the groups and the event probability of the "optimal strategy" were 0.75 for the GPs and the cardiologists, and 0.68 for the students.

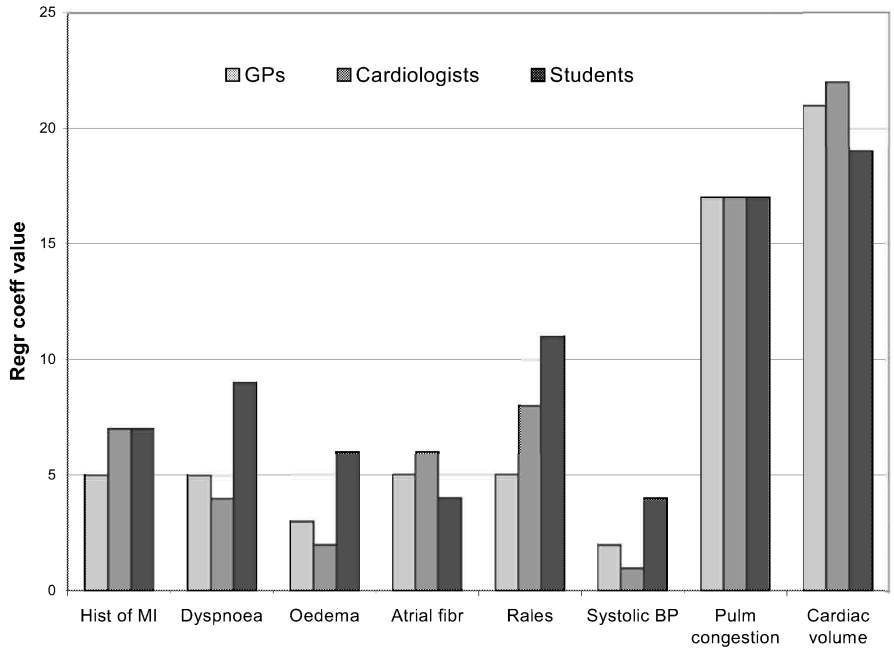


Figure 7. The GPs', cardiologists', and students' judgements strategies.

The adjusted coefficient of determination (R^2 adj) was 68 for the GPs, 66 for the cardiologists and 70 for the students.

Diagnostic accomplishment

For each individual case vignette, the variation between the probability assessments made by the different participants was large. The differences between the minimum and the maximum assessments for individual vignettes varied between 48 and 100 percentage units. The three groups of participants did not differ in this respect.

We measured the diagnostic accomplishment as the difference between the mean probability assessments for the groups of patients with and without CHF. The best result that could be expected was 55 percentage units (the difference between the average event probability for vignettes with and without CHF, respectively, see page 23). The figures for the GPs, the cardiologists and the students were 20, 21, and 20 percentage units, respectively, and for the three clusters 24, 18, and 19 percentage units, respectively. Thus the GPs, cardiologists and students did not differ in this respect, but the first cluster strategy, in which cardiac volume was the most important cue, turned out to be better than the other two cluster strategies.

Cardiac enlargement was the most important cue in the optimal strategy, while the relative importance of the rest of the cues could not be unequivocally assessed owing to the small number of case vignettes. However, several combinations of cues may work

equally well for the judgements, and different subsets may be utilised, depending on the information available. For that reason, we tested models based on various combinations of cues for their sensitivity and specificity. We also looked at individual doctors' models, and their diagnostic accomplishment (Table 5).

Table 5. Sensitivity and specificity of different models and strategies.

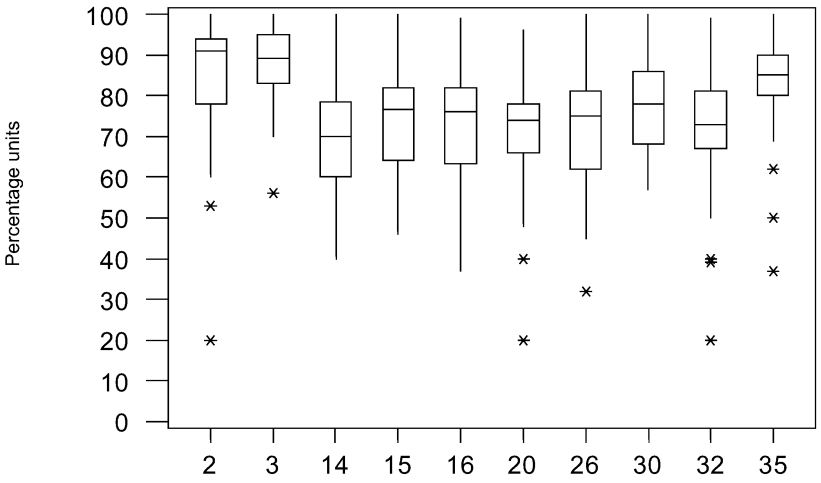
Model	Sensitivity	Specificity
Computer models		
Cardiac enlargement only	0.77	0.71
Cardiac enlargement and pulmonary stasis	0.88	0.64
Classical symptoms and signs (dyspnoea, oedema, rales)	0.62	0.79
History and examination (all except lung and heart X-ray)	0.85	0.79
"Optimal strategy" (computer model with 8 cues)	0.92	0.86
Participants' judgement strategies		
Cluster 1 (cardiac enlargement the most important cue)	0.69	0.86
Cluster 2 (pulmonary congestion the most important cue)	0.58	0.64
Cluster 3 (more even weights of the cues)	0.65	0.79
Doctor A (participant with the highest accomplishment)	0.81	0.79
Doctor B (participant with the lowest accomplishment)	0.50	0.64

STUDY III

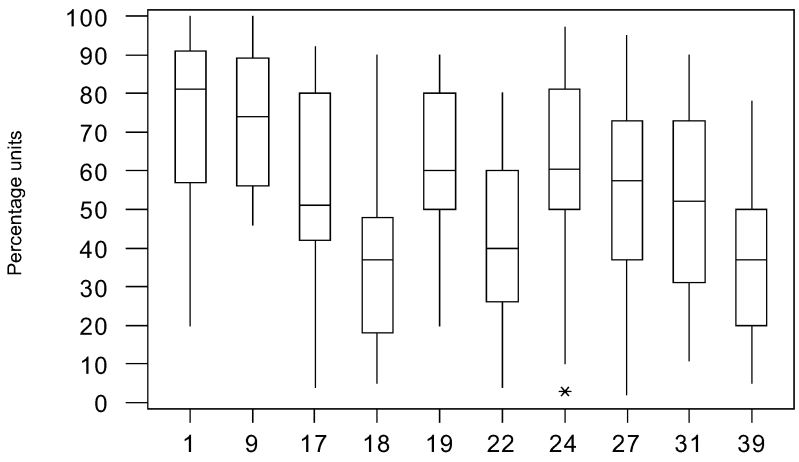
In this study, patient characteristics were investigated and related to the participants' agreement or disagreement regarding the diagnosis.

The 10 case vignettes with the least and the 10 with the most diverging assessments were compared to each other. They differed in assessed probability of CHF (Figure 8). The only significant cue differences between the two groups concerned cardiac enlargement and atrial fibrillation (Table 6). Some cue combinations were also investigated (cardiac enlargement & a history of myocardial infarction; cardiac enlargement & dyspnoea; cardiac enlargement & dyspnoea & atrial fibrillation), but the groups did not differ in those respects. The "classic" CHF findings, dyspnoea, oedema, and rales, were only found in three of the 40 cases.

The regression line plot indicates that the case vignettes with the least divergent assessments were those with the highest and the lowest number of positive cues, and the case vignettes with the most divergent assessments were those with an intermediate number of cues (Figure 9).



The case vignettes with the least divergent assessments (Heart failure: vignettes 2, 3, 14, 15, 16, 30, 32, and 35)



The case vignettes with the most divergent assessments (Heart failure: vignettes 9, 19, 22, 24, and 31)

Figure 8. The case vignettes with the least and the most divergent assessments. The box size (= the interquartile range) reflects the participants' divergence in rating the probability of heart failure for each individual patient. The bottom of the box is at the first quartile (Q1), the top is at the third quartile (Q3), and the line across the box is at the median value. The "whiskers" (= the lines that extend from the top and bottom of

the box) extend to the smallest and the largest observation (= participant) that is not considered an outlier. Outliers (*) are observations outside these limits.

Table 6. Characteristics of the case vignettes representing the least and the most divergent assessments.

	Least divergent assessments $n = 10$	Most divergent assessments $n = 10$
Number of positive cues	4.5	2.7
Number of patients with heart failure	8	5
Age, mean and range	75 (61-92)	75 (56-84)
Sex, number of men	6	6
Number of patients with a history of myocardial infarction	3	5
Number of patients with dyspnoea	10	7
Number of patients with atrial fibrillation **)	8	1
Number of patients with leg oedema	3	4
Number of patients with rales	5	3
Number of patients with a systolic blood pressure ≤ 140 mmHg	6	4
Number of patients with signs of pulmonary stasis	5	4
Number of patients with cardiac enlargement	10	6
Number of patients with cardiac volume >700 ml/m^2 **)	8	1
Relative cardiac volume (ml/m^2), mean and range ***)	791 (600-920)	533 (370-900)
Echocardiogram: ejection fraction (%), mean and range (not presented as a cue)	40 (25-58)	44.6 (25-55)

**) $p < 0.01$
***) $p < 0.001$

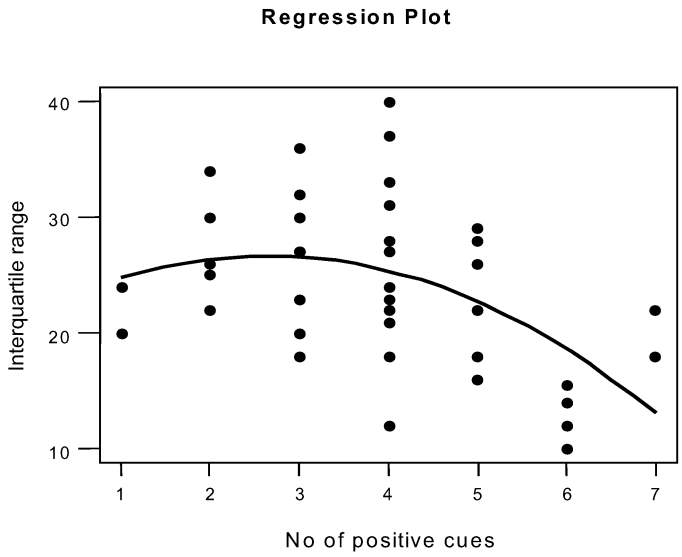


Figure 9. Association between the degree of assessment divergence (interquartile range) and the number of positive cues. A regression line plot representing the association between the interquartile range values and the number of positive cues for the 40 case vignettes.

STUDY IV

In this study, theoretical aspects of creating representative sets of case vignettes for CJA studies were discussed (see Introduction, Representative case vignettes), as well as practical problems. The two sets of case vignettes that were constructed and utilised in Study I and Studies II-III were used to exemplify problems and possible consequences.

Selecting relevant populations and cases

In Study I we considered it important to select cases from primary health care in order to construct a set of case vignettes that would represent PHC patients. We had decided not to present ECHO as a cue in the case vignettes, but we wanted information about it since it is often used as a diagnostic gold standard. However, information about ECHO was scarce in the patient records, and we had to give up this ambition. Information about other relevant types of information was also difficult to find for many of the patients, and this turned out to be an important limiting factor for selecting cases for our study. In order to have a better chance of finding the information we wanted, we had to select patients who had been diagnosed as having CHF. However, the validity of CHF diagnoses in PHC is known to be low, and the probability of CHF among this group of patients might have been 50-70% (Remes et al., 1991; Wheeldon et al., 1993).

In Study II we wanted to compare the diagnostic accomplishment of GPs, cardiologists, and medical students, and we therefore wanted to include cases with gold standard diagnoses (CHF and non-CHF). We also wanted the cases to be representative of PHC. We therefore selected our cases among a group of patients referred by GPs to a cardiology department during a three-year period. The cases thus represent a selected group of PHC patients. In this group, too, missing information in the patient records turned out to be an important limiting factor when selecting cases. The probability of CHF in the group of cases was high, 65%.

In Table 7 the characteristics of the two sets of case vignettes are presented. The two sets represent different populations, but they are still surprisingly similar. The cases in Study I were older, and fewer of them were men, but no significant differences ($p<0.05$) were found between the two sets of case vignettes regarding the seven variables they had in common.

Table 7. Comparison of the two sets of case vignettes (the duplicate cases in Study I are excluded)

Common variables	S1 $n = 40$ (%)	S2 $n = 40$ (%)
History of myocardial infarction	8 (20)	12 (30)
Dyspnoea	31 (78)	32 (80)
Oedema	20 (50)	13 (33)
Atrial flutter/irregular heart rhythm	14 (35)	17 (43)
Rales	14 (35)	12 (30)
Cardiac enlargement on X-ray	27 (68)	33 (83)
Pulmonary stasis on X-ray	21 (53)	16 (40)
General characteristics		
Sex: male*	13 (33)	22 (55)
Age number ≥ 80 years old*	25 (63)	14 (35)
Age mean (range)	78 (60-88)	76 (56-92)

*) p-value < 0.05 (chi-square analysis)

Relevant cues, number of cues, number of cases

Textbooks, guidelines, scientific articles, and interviews with GPs about what information they considered to be important guided our decisions about what kind of information we should include as cues in the case vignettes. However, information that seldom or never was found in the patient records could not be included, even if considered important, as for example gallop rhythm and neck vein distension (Davie et al., 1997). ECHO was not used in the case vignettes. The main reason for this was that ECHO is often considered a gold standard for CHF diagnosis, and we therefore expected it to dominate other types of information, since in CJA, all information regarding a case vignette is presented at the same time. Results from Study I also guided our selection of cues for Study II.

Since the number of cues is one of the factors that determine the number of case vignettes, we had to take into consideration how much patience we could expect from our participants.

Informing participants about the selection of patients

The participants in our studies were informed that the case vignettes were based on authentic patients who had been seen by GPs, but they were not told about the selection process in detail.

Some general remarks

Conducting a representative CJA study in clinical medicine in accordance with theoretical demands is difficult, and involves extensive work. It will probably never be possible to fulfil all the requirements, and the problem will rather be to make compromises that will not have serious consequences regarding the validity of the study. In our two studies, the choice of variables and the quality of documentation in the patient records turned out to be decisive for the process of selecting patients for our patient descriptions. The setting (primary health care with a variety of diagnostic judgements) and the task itself (heart failure, which is a chronic disease) added to the difficulties.

An alternative to collecting authentic patients could have been to create a set of "artificially representative case vignettes". However, given the scarcity of knowledge in primary health care about the characteristics of heart failure patients, this would have been a risky project. A wide range of patients might have been acceptable to the participants, and even if the set of case vignettes and the distribution of characteristics had not represented the group of relevant patients in a statistical sense, they might have served our purpose. Using authentic patients, however, guaranteed that patients like those we presented in the case vignettes really existed. We could also tell the participants that the patients were authentic, and when asked, they reported that the patients could have been taken from their own group of patients. Since the patients in Study II were thoroughly investigated, we could also use the diagnoses as a gold standard for evaluating the diagnostic accomplishment.

STUDY V

Participants, response rate

Thirty GPs were contacted and 15 of them agreed to participate. The participants were on average 52.7 (range 42-62) years of age, they had been specialists in family medicine for an average of 14.8 (range 3-25) years, and six of them were men. The response rate was 50%. Those who declined to participate were not asked why they did so, but many said spontaneously that the reason was a heavy workload. The non-participating GPs were on average also 52.7 (range 35-62) years of age, and seven of them were men.

Diagnostic reasoning

The information that was used most frequently in diagnostic arguments was the information about the ejection fraction value on ECHO, pulmonary congestion, and cardiac volume (Figure 10). The most frequent argument for CHF was pulmonary

congestion, and the most frequent argument against CHF was the ejection fraction value. Three quarters of the arguments were in favour of the diagnosis of CHF.

Symptoms and signs were used less frequently as arguments (Figure 10). Symptoms were used most frequently when reasoning about the prototypical CHF case, CV2 (case vignette number two; severe dyspnoea, orthopnea), and the prototypical non-CHF case, CV5 (no dyspnoea). Signs were used most frequently when reasoning about CV1 (rales) and CV3 (tachycardia).

One of the cases (CV5) had a normal ECG, which opposes the diagnosis of CHF according to the guidelines (Table 2). However, only three of the GPs used this as an argument. Chest X-ray findings were frequently used by the GPs as arguments in their diagnostic reasoning (Figure 10). According to guidelines this kind of information supports the diagnosing of CHF or non-CHF, but it is less important than symptoms and signs (Table 2). The ejection fraction value was used in more than half of the judgement situations. However, in a third of the judgements, no ECHO information was used as an argument, in spite of the fact that it is considered a very strong argument both for and against CHF in the guidelines.

In 70 out of 90 judgement situations the GPs used information about other diseases as arguments. Atrial fibrillation, emphysema, history of myocardial infarction and hypertension were the diagnoses most commonly used in this way. Age was seldom used as an argument; it was used only for the two cases over 80 years of age.

For certain variables, the same value was used by some GPs as an argument for and by others as an argument against CHF. The presence of emphysema was sometimes seen as increasing the risk of CHF, and sometimes as an alternative explanation for symptoms. Diabetes could also be seen as increasing the risk of CHF, or as an alternative explanation for symptoms. For relative cardiac volume, the reasoning could be compatible with GPs using different threshold values in their reasoning. For the six cases, the two lowest values for cardiac volume (460 and 470 ml/m^2) were only used as arguments against CHF, the two highest values (820 and 920 ml/m^2) only as arguments for CHF, but the two intermediate values (520 and 630 ml/m^2) were used as arguments in both directions.

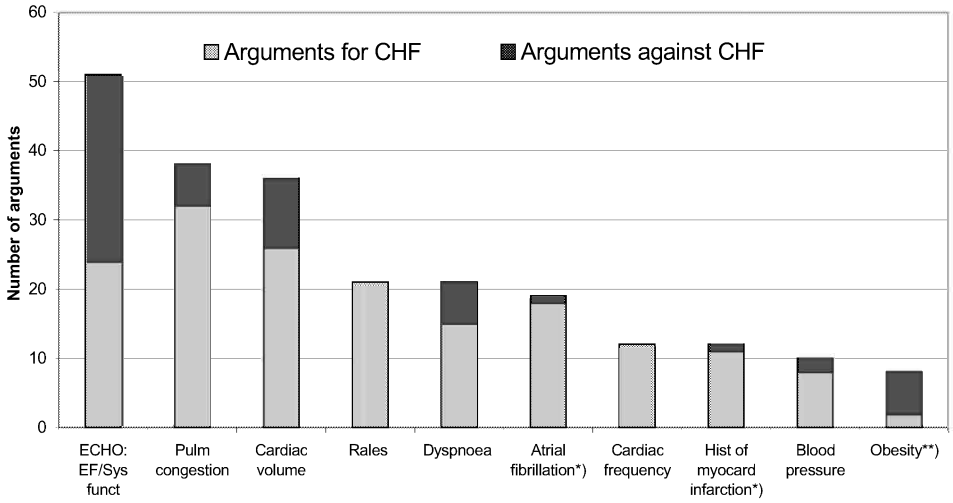


Figure 10. The ten most frequently used arguments, making use of different categories of clinical information. Variables that are not presented in all the vignettes are indicated by * (presented in five vignettes) or by ** (presented in four vignettes).

Diagnostic judgements

There was total agreement among the GPs regarding the diagnosis for the prototypical CHF case, but for the other five cases there was a large variation among the GPs' diagnostic judgements. Case vignettes representing CHF patients were more likely to be correctly diagnosed than those representing non-CHF patients (Table 8).

Reliability – inter-rater coding agreement

Coding of think-aloud protocols: Disagreement regarding the variable content was 4.8% of the tested segments of propositions. The remaining segments were tested for agreement on argument values, which was 95% (κ 0.85).

Classification of diagnostic judgements: The inter-rater agreement was 92% (κ 0.85).

The coding and classification reliability was thus good.

Table 8. Diagnoses of the case vignettes according to the clinical diagnoses and according to the guidelines, the GPs' classification of the case vignettes, and the number of arguments used by the group of GPs. Shadowed cells indicate correct judgements.

	CV2	CV6	CV3	CV5	CV4	CV1
Diagnosis according to cardiologists	CHF	CHF	CHF	Not CHF	Not CHF	Not CHF
Diagnosis according to European guidelines	CHF	CHF	Not CHF?	Not CHF	Not CHF	Not CHF
Total number of arguments used by the GPs (proportion of arguments for CHF)	60 (98%)	52 (75%)	54 (85%)	57 (44%)	63 (70%)	63 (75%)
Number of GPs (*n* = 15) classifying the patient as:						
CHF	15	11	11	3	6	11
Not CHF	0	2	2	9	6	3
Uncertain, no classification	0	2	2	3	3	1
Correct diagnoses of CHF (*n* = 45) and not CHF (n = 45) judgements (proportion of correct diagnoses)	37/45 (CHF 82%)			18/45 (Not CHF 40%)		
Correct diagnoses of all judgements (*n* = 90) (proportion of correct diagnoses)	55/90 (All cases 61%)					

DISCUSSION

The purpose of the studies was to increase our understanding of GPs' diagnostic judgements of patients with CHF. Two different methodological approaches were used for the investigation of the diagnostic judgements, CJA in Studies I, II and III, and think-aloud technique in Study V. These two methods can elucidate in two different ways how clinical information influences the judgements. In Study IV we also analysed the methodological problem of selecting a set of representative patients for CJA studies.

INFORMATION AFFECTING THE JUDGEMENTS

Cue importance

In a wider context, we can say that a cue is important because it has a high sensitivity, a high specificity, a high predictive value in a certain population, or because it is reliable, often available, or perhaps because it represents non-expensive information. In general, cues with high sensitivity and low specificity (e.g. dyspnoea) are more important in an early stage of the diagnostic process, and cues with high specificity and low sensitivity (e.g. gallop rhythm) are more important in a later stage, for validating the diagnosis.

When we say, within the framework of CJA, that a high regression weight means that a certain cue is important for a phenomenon, this can be given different interpretations. It might mean that the cue causes the phenomenon, or that the cue is caused by the phenomenon, or that the cue is an indicator for the phenomenon because they are associated for some other reason (e.g. they have a common cause).

The calculation of regression weights for the individual participants does not mean that they really weigh the cues in every case, i.e. that they use compensatory decision rules. The regression weights represent an overall relation between the cues (the input) and the judgements (the output), but they do not tell us if, for example, 'fast and frugal' decision rules are used, or if the value of one cue possibly changes the value of another cue.

In Studies I-III, all information about one case was presented on one paper, and the doctors thus had information about all the cues when they made their probability assessments. Since information about ECHO often has functioned as a gold standard for diagnosing CHF (Davie et al., 1997; Wheeldon et al., 1993), and therefore might dominate the judgements, we did not include this information in the vignettes, and let the vignettes instead represent an earlier stage in the judgement process. In Study V the information was presented successively in five steps, and information about ECHO on the fifth screen thus was not to influence the GPs' interpretation of the information given on the first four screens.

In Study V, the use of a cue as an argument for or against the diagnosis of CHF was used for measuring its importance. For each individual GP the use of a specific cue was only registered once per case vignette. For a specific cue, then, its importance for

an individual GP could be measured as the number of cases in which the cue was used (0-6), and its importance for the whole group of GPs could be measured as the number of judgement situations in which it was used (0-90).

Since importance thus has different interpretations in the CJA and the think-aloud studies, comparisons across the studies should be made with caution.

Cues of importance for the judgements

Diagnostic judgements were made by GPs in Studies I, II, and V. In both Study I and Study II, relative cardiac volume and pulmonary congestion were the two most important cues, both measured by the regression coefficients on a group level and by the number of participants utilising them as the most important cue. The value of the relative cardiac volume cue was dichotomous in Study I, while in Study II it was continuous. The regression coefficient in Study I thus reflected the change when going from no cardiomegaly to cardiomegaly, while in Study II it reflected the change when cardiac volume increased by 300 ml/m^2. However, this does not change the fact that these two cues were the most important cues in both studies. In Study V we had included ECHO, which turned out to be the most frequently used cue, but the two second most frequently used cues in this study were also relative cardiac volume and pulmonary congestion.

Although ECHO was the most frequently utilised cue in Study V, it received no attention in one third of the judgement situations, and this finding might indicate that it could have been included among the cues in Studies I and II without distorting the results.

In Study II, GPs were compared with cardiologists and medical students. No differences on the group level were found. This could be an effect of the method, which may not do justice to important aspects of expert competence (such as perceptual skill, for example) or it could be an effect of the selection of case vignettes (too paradigmatic?).

Linear models may not do full justice to all the participants. Some of them may use interactions between cues or non-linear models, and their strategies might therefore be underestimated. However, linear regression models are robust, and the fit in both studies was good for most of the participants.

GOLD STANDARDS FOR DIAGNOSTIC JUDGEMENTS AND REASONING

Study I had an explorative character, and given the aim (to study cue utilisation and variation), a gold standard diagnosis for each case was not necessary. In Studies II and III, one of the aims was to evaluate the participants' diagnostic accomplishment, and a gold standard for the CHF diagnosis was therefore necessary. In Study V, the aim was to study both the diagnostic judgements and the diagnostic reasoning, and a gold standard for both the CHF diagnosis and the reasoning process was needed.

In Studies II-III, the clinical diagnoses made by cardiologists at the cardiology department were considered to be the best obtainable gold standard. However, because the diagnoses were expressed as CHF or not CHF, they could not be used for direct comparison with the probability assessments made by the participants. They were used for calculation of an "optimal model", i.e. the regression weights on the left side of the lens model (Figure 4). This, in turn, was used for calculating the average probability of CHF for the group of case vignettes with and without the diagnosis CHF, respectively. This difference was used as a measurement of the possibility to discriminate between CHF and non-CHF. The regression weights in the "optimal model" represent a change in odds and were therefore not suitable for direct comparisons with the participants' regression weights, which represent a change in probability.

A problem with the "optimal model" is that it has been calculated to exactly fit the group of cases utilised in Study II. The representativity of the model is therefore dependent on the representativity of the group of case vignettes, and it would be unfair to the participants to use it as a gold standard if the group of cases differed very much from the cases they usually meet. However, since relative cardiac volume was the dominating cue in the "optimal model" and this was in agreement with epidemiological studies in PHC, this might indicate that the model could be acceptable (Davie et al., 1997).

In Study V, the case vignettes were selected from the Study II cases, and the clinical diagnoses could therefore be used. The way the participants' diagnostic judgements were classified (CHF, not CHF, no diagnosis) made it possible to match them directly with the clinical diagnoses. For the diagnostic reasoning the European guidelines were chosen as a gold standard (Remme et al., 2002; The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology, 2001). However, since the guidelines mainly consist of reasoning text, they are difficult to operationalise, and therefore the table of assessments to be performed routinely to establish the presence of CHF (Table 2) was chosen as gold standard. The two necessary conditions for CHF in the definition (Table 1) can be said to be included in this table. Text-based guidelines have been shown to be more difficult to use in daily practice than guidelines in an algorithm format (Patel et al., 2001).

The different assessments (types of information) in the table (Table 2) are weighted with (+) or (+++), according to the support they are supposed to give to the diagnoses CHF or non-CHF, and this weighting can in some sense be interpreted as the "importance" of the different types of information. However, it is by no means obvious if and how this importance can be seen as a norm for how frequently these types of information ought to be used by the participants when judging the case vignettes. Nor is it obvious how information should be weighed together – by some compensatory decision rule, or by some other kind of decision rule. When the two necessary conditions in the definition were tested as a gold standard for diagnosing the case vignettes, this did not give correct results for the complex case vignette CV3 (Table 8), indicating that the definition might not be adequate in all judgement situations. Would it be possible, for example, that strong support from two or more assessments could make up for the lack of a necessary condition?

DIAGNOSTIC JUDGEMENTS

A number of studies have showed that the diagnosis CHF in PHC may be wrong for 30-50% of the patients, and that there is a tendency to over-diagnose CHF (Remes et al., 1991). In Study II, the method used for assessing the diagnostic accomplishment did not allow us to get the number of correct and incorrect diagnoses. However, since the capacity for discrimination between CHF and non-CHF was low, we may infer a rather low diagnostic accuracy. Study III showed that a larger number of cues indicative of CHF tended to make the participants agree more often on the diagnosis. However, this was no guarantee that the diagnosis was correct. In Study V, the GPs correctly diagnosed 61% of the diagnostic situations, and there was a strong tendency to over-diagnose CHF.

When CHF is defined as in the guidelines (Table 1), what is really defined is systolic heart failure. When epidemiological studies of CHF in PHC use the ejection fraction as gold standard for the correct diagnosis, they are dealing with systolic heart failure. However, there is an increasing interest in diastolic heart failure, which is more difficult to describe and define (Caruana et al., 2000; European Study Group on Diastolic Heart Failure, 1998; Spencer & Lang, 1997; Zile, 2003; Zile & Brutsaert, 2002). The recently suggested definition of diastolic heart failure (symptoms and signs of heart failure and LVEF > 50%) (Zile & Brutsaert, 2002) makes it difficult to state that CHF in a wider sense is really over-diagnosed in PHC. It also creates a problem for our studies, since we did not ask the participants specifically to assess the probability of systolic heart failure (Studies I-III) or to try to decide whether the patients had systolic heart failure or not (Study V), and it might therefore be difficult to do justice to their answers.

METHODOLOGICAL CONSIDERATIONS

Participants in the studies

The conclusions from the study, and in particular the comparisons between the studies, should be regarded with caution because of the size of the non-responding groups. In Study I, the group of GPs were not selected at random, and no response rate could be calculated, and the group might therefore not be representative for GPs.

In Studies II-III, the response rates for both GPs and cardiologists were about 55%, and the representativity may therefore be less good. Cardiologists from two hospitals were invited, but the participants in the study came mainly from one of them (one of the authors worked at that hospital). If there were a systematic difference between the judgement policies at the two hospitals, this might thus have influenced the results. We had expected the students to have a low response rate, and since we wanted the three groups to be of about the same size, we invited a large group of students to participate. Because of their potential dependency situation, we had promised the students full anonymity, and therefore no follow-up of non-participants was possible. The participating students might well have been more interested in cardiology or more self-confident than the non-participants, and this might have influenced the results.

In Study V, the response rate was rather low, which might influence the interpretation of the results. However, the age and sex distribution were the same for the non-responders as for the responders.

Validity of case vignettes

The quality of the conclusions from the study depends on the assumption that decisions about case vignettes reflect decisions about real patients, i.e. the validity of case vignettes. Case vignettes have been used in a number of studies since they represent a convenient way of studying clinical problems: different doctors can judge the same cases, the information can be presented in a well-defined way, and the information content can be varied systematically. However, few studies have addressed the validity.

In a systematic review of 74 articles, only 11 discussed the validity problem, and of those, only two were designed so that a comparison between clinical behaviour and behaviour with the case vignettes could be assessed (Jones et al., 1990). In these two articles there was a strong correlation between rheumatologists' judgements of disease activity in their own patients and their judgements of disease activity in case vignettes based on these patients, and this could also be reproduced with another group of rheumatologists ($r = 0.90$, and $r = 0.85$, respectively) (Kirwan et al., 1983a; Kirwan et al., 1983c). The validity of case vignettes was also validated for GPs assessing patients with mental health problems (Braspenning & Sergeant, 1994). In another study, the doctors' judgement policies were found to be stable when reinvestigated after a year (Kirwan & Currey, 1984). This supports the idea that case vignettes can have high validity.

In the present studies, the validity was not evaluated. In Study I, however, 24 out of 27 GPs answered yes when asked whether the cases could have been taken from their own health centre, and this gives some support to the face validity of the case vignettes. The use of authentic patients for the construction of case vignettes should also contribute to the validity of the cases. However, case vignettes could be made more realistic if, for example, photographs of "patients" could be added.

CJA

In many CJA studies, standardised beta coefficients are used as a measurement of cue importance. In our studies we decided to use regression coefficients, since regression coefficients are easy to interpret and can be adapted to different types of cues by use of clinically relevant scale steps.

In a typical CJA study the case vignettes contain information about all the cues. This might be more suitable for fairly simple on-the-spot judgements, like prescribing an antibiotic for a child with tonsillitis, than for chronic diseases like CHF. CHF is often not diagnosed during a single encounter, but rather over a series of encounters during which more and more information is collected. This may make the judgement situation in Studies I-III less realistic. However, since hypotheses are formed early in clinical encounters (Elstein et al., 1978), it might not be unreasonable to expect the participants to think about CHF, given the case vignettes they were presented with.

Another problem with this way of presenting a case is that a very specific cue might predominate over the other cues, and the supposed integration of all cues might not take place. The necessity to present information about the same cues in all the vignettes may also be problematic. If every case vignette contained information about a cue that is seldom used by the participants, this could have a reminding effect ("cuing effect") that could influence the results. Giving information to GPs about gallop rhythm in every case vignette could, for example, have this effect.

A typical CJA study also means that the participants are presented with a fairly large number of similar looking case vignettes, which makes the task seem very repetitive. This could be expected to lead to boredom and negligence at the end of the task. However, both the rather high value of adjusted R^2 , and comparisons we made between the participants' models for the first and the last 20 cases, respectively, would speak against this effect.

Think-aloud

The characteristics of the case vignettes will influence which information is used as arguments. All our six cases had different degrees of rales, and this may explain why information about rales was used quite often as an argument, but also that it was only used as an argument for CHF. Absence of rales could, for example, have been used an argument against CHF. Some of the GPs also commented upon the frequency of rales among the cases. Only one of the cases had a normal ECG, and thus there was only a limited number of judgement situations that allowed the GPs to make use of this information. The same was true of gallop rhythm, which also was present in only one case. However, out of the 15 GPs, only three used the normal ECG and one the gallop rhythm as arguments.

Case characteristics in a more general sense may also influence the results (Svenson, 1989). If making the judgements is very much of a routine task, the strategies may be highly automatised and the degree of verbalisation low, which will result in poor verbal protocols. If there are many cases of the same kind, this may involve a learning process, which could contribute to some degree of automatisation toward the end of the task. Judgements that are considered important often tend to give richer verbal protocols. Time spent on a case may be correlated to how difficult it is (Svenson, 1979). Our cases were rather complex and contained much information, and the average time the GPs spent on a case varied from two to seven minutes. The relation between the length of the verbal protocols, the judgement decisions, and the reasoning process will be further investigated in a forthcoming study.

The GPs sometimes mentioned lack of certain information, such as liver tests, follow-up of diabetes tests, and exercise tests. Different GPs seemed to systematically lack different types of information. This was not analysed within the scope of the present study.

Some of the GPs sometimes discussed an issue while using the pointer to indicate on the screen what they were talking about. This creates a methodological problem,

since in the verbal protocols an issue might sometimes have been commented upon that could not be identified.

When working with the cases, the GPs often tended to become very involved in the "patients". They did not limit themselves to the task, which was to reason about the diagnosis, but discussed treatments, referrals, and nursing activities as well. Some of them could become quite "emotional" about some of the patients, saying things like "Oh, this poor man, really in bad shape…" or "Ooh, she's so sick!" This might be taken as an indicator of the face validity of the case vignettes and the task.

Benefit of the studies

The studies in this thesis are examples of interdisciplinary research. Clinical medicine has supplied us with the research problems, and cognitive psychology with the research methods. This combination is of great interest, but until now little of this type of research has been done in Sweden, and the experiences from these studies will be of great use in forthcoming investigations.

Electronic patient records are becoming more common in health care. We are no longer satisfied with using them as advanced typewriters, but want them to help us to present and structure data in a way that will aid us in our decisions, and make the growing body of medical knowledge easy to access in our daily practice. In this work, and in the work with various types of decision support systems or guidelines, there will be greater need for knowledge about our cognitive processes (Karlsson, 2001; Kushniruk et al., 1995; Kushniruk et al., 1996; Kushniruk & Patel, 1998; Kushniruk et al., 1998; Patel et al., 2001; Patel et al., 1985; Patel et al., 2000).

CONCLUSIONS

There was large variation in the diagnostic strategies of the participants, measured as the patterns of regression weights for the cues. However, they were consistent in their strategies. On the group level, no differences were found between GPs, cardiologists and medical students regarding their strategies. The most important cues in the participants' strategies were cardiac volume and pulmonary congestion. With cluster analysis three different strategies were identified, one dominated by cardiac volume, one by pulmonary congestion, and one characterised by a more even utilisation of the cues. When giving their own opinion about their diagnostic strategies, the GPs over-estimated the importance of dyspnoea.

There was also large variation in the diagnostic judgements of the participants, measured as the assessed probability of CHF. On the group level, no differences were found between GPs, cardiologists and medical students regarding their diagnostic accomplishment. The participants who used the strategy dominated by cardiac volume had a better diagnostic accomplishment than the others, but they were still far from the optimal accomplishment, given the information in the case vignettes.

For each case vignette the diagnostic judgements varied widely. A large number of cues indicative of heart failure made it easier for the participants to agree about the diagnostic judgements. Increased cardiac volume and presence of atrial fibrillation also increased the diagnostic agreement between participants.

When we constructed the sets of representative case vignettes for the CJA studies, two factors above all turned out to be problematic: the incomplete information in the patient records and the necessity of keeping the number of case vignettes low. These two factors necessitated compromises regarding, for example, the choice of cues and the number of cues.

The GPs' use of the clinical information in the case vignettes was not optimal if compared with the recommendations in the European guidelines. Information about ECHO was the information that was used most frequently as a diagnostic argument by the GPs, and this was in congruence with guidelines. However, ECHO seemed not to have been taken into consideration in a third of the judgement situations, which is remarkable. Information about other diseases was frequently used by the GPs in their diagnostic reasoning, but this is not reflected in the guidelines.

FORTHCOMING RESEARCH

The verbal protocols from the think-aloud study will be further analysed with respect to the reasoning process. For example, temporal aspects regarding the reasoning, use of decision rules, differences in reasoning patterns between GPs ending up diagnosing a case as CHF or non-CHF, will be considered. The GPs also answered the question ”What do you consider most important when diagnosing heart failure?” and these protocols will be analysed.

We plan to do a think-aloud study with the same case vignettes, but with cardiologists as participants. Do cardiologists use other cues or other decision rules? Do they use different kinds of knowledge? Do they reason differently regarding the ECHO results? These findings can be of interest in constructing guidelines and decision support systems.

It would also be of interest to re-analyse data from Study II using the Fast & Frugal paradigm. If we find simpler models, possibly more similar to the doctors' judgements in real life, this might be constitute better input to decision support systems.

ACKNOWLEDGEMENTS

This thesis is not just one person's work. It has been possible through co-operation and collaboration and with the support of many people. In particular I wish to express my gratitude to:

Associate Professor *Lars-Erik Strender*, my supervisor, for constant interest and support, and for wisely linking other experts to my thesis project, which has contributed to the quality of the work.

Associate *Professor Johan Bring*, my associate supervisor, for bringing statistical expertise and intellectual creativity into the work.

Professor *Jan Sundqvist* and Professor Emeritus *Hans Åberg*, for giving me the opportunity to do my thesis work at Family Medicine Stockholm.

Professor *Henry Montgomery*, for generously sharing his expertise in cognitive psychology with me, and for introducing me to his international network for research on decision-making.

Lars Backlund, for collaboration and for many fruitful discussions.

Bengt Ullman, for collaboration and for bringing cardiology expertise into the thesis work.

Professor *James Shanteau*, for helping me see some of my data in a new light.

Jan Odelstad, my supervisor in philosophy, for introducing me to the world of decision theory and for making the difficult work of paper writing enjoyable by means of his support and enthusiasm.

Linnéa Löwgren, for transcribing the protocols.

Jane Wigertz, for revising the English text.

Wilhelm Montgomery, for help with the software QA™ for computer presentation of case vignettes.

The whole staff and fellow research students at Family Medicine Stockholm, for their interest and for many fruitful discussions throughout the years.

Participants in the Network for Medical Decision-Making and Informatics at Family Medicine Stockholm, for supplying an arena for discussions that could lead to new projects.

Co-workers and heads of the different departments where I have worked while carrying out the studies in this thesis, for their patience and support.

All GPs, cardiologists and students who participated in the studies, for taking their time to share their diagnostic reasoning with me.

My family and my friends, for being here with me.

This thesis was supported by grants from the Stockholm County Council and the Swedish Heart Lung Foundation.

REFERENCES

Andre M, Borgquist L., Foldevi M, & Molstad S. (2002). Asking for 'rules of thumb': a way to discover tacit knowledge in general practice. *Fam Pract, 19*(6), 17-22.

Andre M, Borgquist L, & Sigvard M. (2003). Use of rules of thumb in the consultation in general practice--an act of balance between the individual and the general perspective. *Fam Pract, 20*(5), 514-519.

Aucott J N, Taylor A L, Wright J T J et al. (1994). Developing Guidelines for Local Use: Algorithms for Cost-Efficient Outpatient Management of Cardiovascular Disorders in a VA Medical Center. *J Qual Improv, 20*(1), 17-32.

Backlund L, Danielsson B, Bring J, & Strender L-E (2000). Factors influencing GPs' decisions on the treatment of hypercholesterolaemic patients. *Scand J Prim Health Care, 18*(2), 87-93.

Badgett R, Mulrow C, Otto P, & Ramirez G. (1996). How well can the chest radiograph diagnose left ventricular dysfunction? *J Gen Intern Med, 11*, 625-634.

Badgett R G, Lucey C R, & Mulrow C D (1997). Can the clinical examination diagnose left-sided heart failure in adults? *JAMA, 277*(21), 1712-1719.

Bergsten-Rydén T, & Andersson F. (1999). The health care costs of heart failure in Sweden. *J Int Med, 246*, 275-284.

Blankfield R P, Finkelhor R S, Alexander J J et al. (1998). Etiology and diagnosis of bilateral leg edema in primary care. *Am J Med, 105*, 192-197.

Borduas F, Carrier R, Drouin D, Deslauriers D, & Tramblay G. (1998). An interactive workshop: An effective means of integrating the Canadian Cardiovascular Society clinical practice guidelines on congestive heart failure into Canadian family physicians' practice. *Canadian J of Cardiol, 14*(7), 911-916.

Braspenning J, & Sergeant J. (1994). General practitioners' decision making for mental health problems: Outcomes and ecological validity. *J Clin Epidemiol, 47*(12), 1365-1372.

Brehmer A, & Brehmer B. (1988). What have we learned about human judgement from thirty years of policy capturing? In B. Brehmer & C. R. B. Joyce (Eds.), *Human Judgment. The SJT View.* (pp. 75-114). Amsterdam: North Holland Elsevier.

Brehmer B. (1980). In one word: Not from experience. *Acta Psychologica, 45*, 223-241.

Brehmer B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137-154.

Brehmer B, & Joyce C R B (Eds.). (1988). *Human Judgement. The SJT View.* Amsterdam: North-Holland Elsevier.

Brown R L, Brown R L, Saunders L A, Castelaz C A, & Papasouliotis O. (1997). Physicians' Decisions to Prescribe Benzodiazepines for Nervousness and Insomnia. *J Gen Intern Med, 12*, 44-52.

Brunswik E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychol Rev, 62*(3), 193-217.

Brunswik E. (1956). *Perception and the representative design of psychological experiments.* (2nd ed.). Berkely, CA: University of California Press.

Butman S M, Ewy G A, Standen J R, Kern K. B, & Hahn E. (1993). Bedside cardiovascular examination in patients with severe chronic heart failure: importance of rest or inducible jugular venous distension. *J Am Coll Cardiol, 22*, 968-974.

Cabana M D, Rand C S, Powe N R et al. (1999). Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA, 282*(15), 1458-1465.

Carlson K J, Lee D C-S, Goroll A H, Leahy M, & Johnson R A. (1985). An analysis of phycisians' reasons for prescribing long-term digitalis therapy in outpatients. *J Chron Dis, 38*(9), 733-739.

Caruana L, Petrie M C, Davie A P, & McMurray J J V (2000). Do patients with suspected heart failure and preserved left ventricular systolic function suffer from "diastolic heart failure" or from misdiagnosis? A prospective descriptive study. *BMJ, 321*, 215-218.

Chapman G B, & Sonnenberg F A (Eds.). (2000). *Decision Making in Health Care: Theory, Psychology, and Applications.* Cambridge UK: Cambridge University Press.

Chaput de Saintonge D M & Hathaway N R. (1981). Antibiotic use in otitis media: Patient simulations as an aid to audit. *BMJ, 283*, 883-884.

Chaput de Saintonge D M & Hattersley L A. (1985). Antibiotics for Otitis Media: Can We Help Doctors Agree? *Fam Pract, 2*, 205-212.

Charlin B, Tardif J, & Boshuizen H P A. (2000). Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine, 75*, 182-190.

Chi M T H, Glaser R & Farr M J. (Eds.). (1988). *The nature of expertise*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Clarke K W, Gray D & Hampton J R. (1994). Evidence of inadequate investigation and treatment of patients with heart failure. *Br Heart J, 71*, 584-587.

Cleland J G F & Habib F. (1996). Assessment and diagnosis of heart failure. *J Intern Med, 239*, 317-325.

Cline C M J, Boman K, Holst M, Erhardt L R & Swedish Society of Cardiology Working Group for Heart Failure. (2002). The management of heart failure in Sweden. National report. *Eur J Heart Fail, 4*, 373-376.

Coderre S, Mandin H, Harasym P & Fick G. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education, 37*, 695-703.

Connolly T. (1988). Hedge-clipping, tree-felling and the management of ambiguity: the need for new images of decision-making. In L. R. Pondy & R. J. Boland & H. Thomas (Eds.), *Managing ambiguity and change* (pp. 37-50). New York: John Wiley.

Cooksey R W. (1996). *Judgment Analysis. Theory, Methods, and Applications*. San Diego: Academic Press.

Cowie M, Struthers A, Wood D et al. (1997). Value of natriuretic peptides in assessment of patients with possible new heart failure in primary care. *Lancet, 350*(9088), 1349-1353.

Custers E J F M, Boshuizen H P A & Schmidt H G. (1996). The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates. *Memory & Cognition, 24*, 384-399.

Custers E J F M, Boshuizen H P A & Schmidt H G. (1998). The Role of Illness Scripts in the Development of Medical Diagnostic Expertise: Results From an Interview Study. *Cognition and Instruction, 16*, 367-398.

Davidson C. (1996). Investigation in general practice of patients with suspected heart failure [letter; comment]. *Heart, 75*(6), 643.

Davie A, Francis C, Caruana L, Sutherland G & McMurray, J. (1997). Assessing diagnosis in heart failure: which features are any use? *QJM 1997, 90*(5), 335-339.

Davie A P, Francis C M, Love M. P, Caruana L, Starkey I R, Shaw T R D, Sutherland G R & McMurray J J V. (1996). Value of the electrocardiogram in identifying heart failure due to left ventricular systolic dysfunction. *BMJ, 312*(27 January), 222.

Davis D A, Thomson M A, Oxman A D & Haynes R B. (1995). Changing physician performance. A systematic review of the effect of continuing medical education strategies. *JAMA, 274*(9), 700-705.

Davis R C, Hobbs F D R, & Lip G Y H. (2000). ABC of heart failure: History and epidemiology. *BMJ, 320*, 39-42.

Denolin H, Kuhn H P, Krayenbuehl F, Loogen F & Reale A. (1983). The definition of heart failure. *Eur Heart J, 4*, 445-448.

Dhami M K & Harries C. (2001). Fast and frugal versus regression models of human judgement. *Thinking and Reasoning, 7*(1), 5-27.

Eddy D M. (1984). Variations in physician practice: The role of uncertainty. *Health Affairs, 3*, 74-89.

Elstein A S, Holzman G B, Ravitch M M et al. (1986). Comparison of Physicians' Decisions Regarding Estrogen Replacement Therapy for Menopausal Women and Decisions Derived from a Decision Analytic Model. *Am J Med, 80*, 246-258.

Elstein A S, Shulman L & Sprafka S. (1978). *Medical Problem Solving: An analysis of Clinical Reasoning.* Cambridge MA: Harvard University Press.

Ericsson K A & Simon H. (1999). *Protocol Analysis: Verbal reports as data. Revised version.* Cambridge, MA, US: MIT Press.

Ericsson K A & Simon K A. (1984). *Protocol Analysis: Verbal reports as data.* Cambridge, MA, US: The MIT Press.

Ericsson K A & Smith J. (Eds.). (1991). *Toward a General Theory of Expertice*. New York: Cambridge University Press.

Eriksson H, Svärdsudd K, Caidahl, K et al. (1988). Early Heart Failure in the Population (The Study of Men Born in 1913). *Acta Med Scand, 223*, 197-209.

Essex B & Healy M. (1994). Evaluation of a rule base for decision making in general practice. *Br J Gen Pract, 44*, 211-213.

European Study Group on Diastolic Heart Failure. (1998). How to diagnose diastolic heart failure. *Eur Heart J, 19*, 990-1003.

Field M J & Lohr M J. (Eds.). (1990). *Clinical practice guidelines: directions for a new program.* Washington DC: Natinal Academic Press.

Fisher A & Fonteyn M. (1995). An exploration of an innovative methodological approach for examining nurses' heuristic use in clinical practice. *Scholarly Inquiry for Nursing Practice. Vol 9(3) Fal 1995, 263-276., 9*(3), 263-276.

Folland E D, Kriegel B J, Henderson W G, Hammermeister K E & Sethi G K. (1992). Implications of third heart sounds in patients with valvular heart disease. The

Veterans Affairs Cooperative Study on Valvular Heart Disease. *N Engl J Med, 327*, 458-462.

Fonteyn M E, Kuipers B & Grobe S J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research, 3*, 430-441.

Gadsboll N, Hoilund-Carlsen P F, Nielsen G G et al. (1989). Symptoms and signs of heart failure in patients with myocardial infarction: Reproducibility and relationship to chest X-ray, radionuclide ventriculography and right heart catheterization. *Eur Heart J, 10*, 1017-1028.

Gigerenzer G & Goldstein D G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychol Review, 103*(4), 650-669.

González-Vallejo C, Sorum P C, Stewart T R, Chessare J B & Mumpower J L (1998). Physicians' Diagnostic Judgments and treatment decisions for acute otitis media in children. *Med Decis Making, 18*, 149-162.

Green P E. (1974). On the Design of Choice Experiments Involving Multifactor Alternatives. *Journal of Consumer Research, 1*(Septembet), 61-68.

Grimshaw J M & Russell I T. (1993). Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *Lancet, 342*(8883), 1317-1322.

Hammond K R. (1998). Representative design. *Brunswik Society web-site: http://brunswik.org/notes/essay3.html*(september).

Harries C, Evans J S B & Dennis I. (2000). Measuring Doctors' Self-insight into their Treatment Decisions. *Applied Cognitive Psychology, 14*, 455-477.

Harries C, Evans J S B, Dennis I & Dean J. (1996). A clinical judgement analysis of prescribing decisions in general practice. *Le Travail Humain, 59*(1), 87-111.

Henry S B, LeBreck D B & Holzemer W L. (1989). The effect of verbalization of cognitive processes on clinical decision making. *Research in Nursing & Health, 12*(3), 187-193.

Hobbs F, Jones M, Allan T, Wilson S & Tobias R. (2000). European survey of primary care physician perceptions on heart failure diagnosis and management (Euro-HF). *Eur Heart J, 21*(22), 1877-1887.

Houghton A, Sparrow N, Toms E & Cowley A. (1997). Should general practitioners use the electrocardiogram to select patients with suspected heart failure for echocardiography? *Int J Cardiol, 62*(1), 31-36.

Ishmail A, Wing S, Ferguson J, Hutchinson T, Magder S & Flegel K. (1987). Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure. *Chest, 91*, 870-873.

Jackson G, Gibbs C R, Davies M K & Lip G Y H. (2000). ABC of heart failure: Pathophysiology. *BMJ*, 167-170.

Johnson R & Wichern D. (1988). *Applied Multivariate Statistical Methods* (second ed ed.): Prentice.

Jones T V, Gerrity M S & Earp J. (1990). Written case simulations: Do they predict physicians' behavior? *J Clin Epidemiol, 43*(8), 805-815.

Jungerman H. (1983). The two camps on rationality. In R. W. Scholz (Ed.), *Decision making under uncertainty.* Amsterdam: Elsevier.

Kahneman D, Slovic P & Tversky A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Karlsson D. (2001). *Aspects of the use of medical decision-support systems: the role of context in decision support. Linköping studies in science and technology Dissertations ; 662.*, Linköping University, Linköping.

Kee F, Jenkins J, McIlwaine S, Patterson C, Harper S & Shields M. (2003). Fast and Frugal Models of Clinical Judgment in Novise and Expert Physicians. *Med Decis Making, 23*, 293-300.

Khunti K, Hearnshaw H, Baker R & Grimshaw G. (2002). Heart failure in primary care: qualitative study of current management and perceived obstacles to evidence-based diagnosis and management by general practitioners. *Eur J Heart Fail, 4*(6), 771-777.

Kirwan J R, Bellamy N, Condon H, Buchanan W W & Barnes C G. (1983c). Judging "Current Disease Activity" in Rheumatoid Arthritis - An International Comparison. *J Rheumatol, 10*, 901-905.

Kirwan J R, Chaput de Saintonge D M & Joyce C R B. (1990). Clinical Judgement Analysis. *Q J Med, 76*, 935-949.

Kirwan J R, Chaput de Saintonge D M, Joyce C R B & Currey H L F. (1983a). Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of "paper patients". *Ann Rheum Dis, 42*, 644-647.

Kirwan J R, Chaput de Saintonge D M, Joyce C R B & Currey L F. (1983b). Clinical judgement analysis - practical application in rheumatoid arthritis. *Br J Rheum, 22*(supplement), 18-23.

Kirwan J R & Currey H L F. (1984). Clinical judgment in rheumatoid arthritis. IV. Rheumatologists' assessments of disease remain stable over long periods. *Ann Rheum Dis, 43*, 695-697.

Klayman,J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view.* . (pp. 115-162). Amsterdam: North-Holland Elsevier.

Kushniruk A, Patel V & Fleiszer D. (1995). *Analysis of Medical Decision Making: A Cognitive Perspective on Medical Informatics.* Paper presented at the Proc Annu Symp Comput Appl Med Care.

Kushniruk A W, Kaufman D R, Patel V L, Lévesque Y & Lottin P. (1996). Assessment of a computerized patient record system: A cognitive approach to evaluating medical technology. *MD Comput, 13*(5), 406-415.

Kushniruk A W & Patel V L. (1998). Cognitive evaluation of decision making processes and assessment of information technology in medicine. *Int Int J Med Inf, 51*, 83-90.

Kushniruk A W, Patel V L & Marley A A. (1998). Small worlds and medical expertise: implications for medical cognition and knowledge engineering. *International Journal of Medical Informatics, 49*, 255-271.

Lamond D, Crow R A & Chase J. (1996). Judgements and processes in care decisions in acute medical and surgical wards. *J Eval Clin Pract, 2*(3), 211-216.

Leary D E. (1987). From act psychology to probabilistic functionalism: The place of Egon Brunswik in the history of psychology. In M. G. Ash & W. R. Woodward (Eds.), *Psychology in twentieth-century thought and society.* Cambridge, UK: Cambridge University Press.

Lomas J, Anderson G, Domnick-Pierre K, Vayda E, Enkin M & Hannah W. (1989). Do practice guidelines guide practice? The effect of a consensus statement on the practice of physicians. *N Engl J Med, 321*(19), 1306-1311.

Lundborg C, Wahlström R, Diwan V K, Oke T, Mårtenson D & Tomson G. (1999). Combining feedback from simulated cases and prescribing, design and implementation of an educational intervention in primary care in Sweden. *Int J Technol Assess Health Care, 15*, 458-472.

Mair F S., Crowley T S & Bundred P E. (1996). Prevalence, aetiology and management of heart failure in general practice. *Br J Gen Pract, 46*, 55-61.

Malmberg I & Persson U. (2000). Primary health care costs in connection with heart failure surveyed: increased use of ACE inhibitors would be beneficial. (Swedish). *Läkartidningen, 97*(20), 2465-2470.

Marantz P R, Alderman M H & Tobin J N. (1988). Diagnostic Heterogeneity in Clinical Trials for Congestive Heart Failure. *Ann Intern Med, 109*, 55-61.

McClelland G. (1999). Representative and efficient designs. *Brunswik Society web-site: http://brunswik.org/notes/essay5/essay5.html, sept.*

McKee P A, Castelli W P, McNamara P M & Kannel W B. (1971). The natural history of congestive heart failure: The Framingham study. *N Eng J Med, 285*, 1441-1446.

McMurray J J V & Stewart S. (2000). Epidemiology, aetiology, and prognosis of heart failure. *Heart, 83*, 596-602.

Medical Products Agency. (1996). Behandling av akut och kronisk hjärtsvikt rekommendationer (Treatment of acute and chronic heart failure recommendations). *Information från Läkemedelsverket, 7*, 17-56.

Meehl P E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidens*. Minneapolis: University of Minnesota Press.

Mejhert M, Persson H, Edner M & Kahan T. (2001). Epidemiology of heart failure in Sweden - a national survey. *Eur J Heart Failure, 3*, 97-103.

Montgomery W & Hammarberg A. (2000). Question Asker 2000 {Computer software and manual} (Version Retrieved from http://www.dirsoft.com).

Morgan S, Smith H, Simpson I et al. (1999). Prevalence and clinical characteristics of left ventricular dysfunction among elderly patients in general practice setting: cross sectional survey. *BMJ, 318*(7180), 368-372.

Mosterd A, de Bruijne M C, Hoes A W, Deckers J W, Hofman A & Grobbee D E. (1997a). Usefulness of echocardiography in detecting left ventricular dysfunction in population-based studies (The Rotterdam Study). *Am J Cardiol, 179*(1), 103-104.

Mosterd A, Deckers J W, Hoes A W et al. (1997b). Classification of heart failure in population based research: An assessment of six heart failure scores. *Eur J Epidemiol, 13*, 491-502.

Nielsen L S, Svanegaard J, Wiggers P & Egeblad H. (2001). The yield of a diagnostic hospital dyspnoea clinic for the primary health care section. *J Intern Med, 250*(4), 422-428.

Nilsson G & Strender L-E. (2002). Management of heart failure in primary health care. A retrospective study on electronic patient records in a registered population. *Scand J Prim Health Care, 20*, 161-165.

Nisbett R E & Wilson T D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Owen A & Cox S. (2001). Diagnosis of heart failure in elderly patients in primary care. *Eur J Heart Failure, 3*, 79-81.

Owens D K & Nease R F. (1997). A Normative Analytic Framework for Development of Practice Guidelines for Specific Clinical Populations. *Med Decis Making, 17*, 409-426.

Parameshwar J, Shackell M M, Richardson A, Poole-Wilson P A & Sutton G C. (1992). Prevalence of heart failure in three general practices in north west London. *Br J Gen Pract, 42*, 287-289.

Patel V L, Arocha J F, Diermeier M, How J & Mottur-Pilson C. (2001). Cognitive psychological studies of representation and use of clinical practice guidelines. *Int J Med Informatics, 63*, 147-167.

Patel V L, Kaufman D R, Arocha J A & Kushniruk A. W. (1985). Bridging theory and practice: cognitive science and medical informatics. *Medinfo, 8*, 1278-1282.

Patel V L, Kushniruk A W, Yang S & Yale J.-F. (2000). Impact of a computer-based patient record system on data collection, knowledge organization and reasoning. *J Am Med Inform Assoc, 7*, 569-585.

Persson H, Malmqvist K, Carlsson A, Rehnqvist,N & Lundman T. (1994). More and more persons suffer from heart failure. Better treatment and information are necessary. *Läkartidningen, 91*, 3251-3254.

Remes J, Miettinen H, Reunanen A & Pyörälä K. (1991). Validity of clinical diagnosis of heart failure in primary health care. *Eur Heart J, 12*, 315-321.

Remme W, Swedberg K & European Society of Cardiology. (2002). Comprehensive guidelines for the diagnosis and treatment of chronic heart failure. Task force for the diagnosis and treatment of chronic heart failure of the European Society of Cardiology. *Eur J Heart Fail, 4*(1), 11-22.

Russo J E, Johnson E J & Stephens D L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*(6), 759-769.

Schmidt H G & Boshuizen H P A (1993). On Acquiring Expertise in Medicine. *Educational Psychology Review, 5*, 205-221.

Schmidt H G, Norman G R & Boshuizen H P A. (1990). A Cognitive Perspective on Medical Expertise: Theory and Implications. *Acad Med, 65*, 611-621.

Simon H A. (1982). *Models of Bounded Rationality*. Cambridge, Mass.: MIT Press.

Smith J, Bruusgaard D, Bodd E & Hall C. (2001). Relations between medical history, clinical findings and plasma N-terminal proatrial natriuretic peptide in patients in primary health care. *Eur J Heart Fail, 3*(3), 307-313.

Spencer K T & Lang R M. (1997). Diastolic heart failure; What primary care physicians need to know. *Postgrad Med, 101*(1), 63-78.

Spiteri M, Cook D & Clarke S. (1988). Reliability of eliciting physical signs in examination of the chest. *Lancet, 1*, 873-875.

Stevenson L W & Perloff J K. (1989). The limited reliability of physical signs for estimating hemodynamics in chronic heart failure. *JAMA, 261*, 884-888.

Struthers A D. (2000). The diagnosis of heart failure. *Heart, 84*, 334-338.

Svenson O. (1979). Process Descriptions of Decision Making. *Organizational Behavior and Human Performance, 23*, 86-112.

Svenson O. (1989). Eliciting and analysing verbal protocols in process studies of judgement and decision making. In H. Montgomery & O. Svenson (Eds.), *Process and structure in human decision making* (pp. 65-81). Chichester: Wiley.

Tape T G, Heckerling P S, Ornato J P & Wigton R S. (1991). Use of clinical judgment analysis to explain regional variations in physicians' accuracies in diagnosing pneumonia. *Med Decis Making, 11*, 189-197.

Tape T G, Kripal J & Wigton R S. (1992). Comparing methods of learning clinical prediction from case simulations. *Med Decis Making, 12*, 213-221.

The Task Force for the Diagnosis and Treatment of Chronic Heart Failure of the European Society of Cardiology. (2001). Guidelines for the diagnosis and Treatment of Chronic Heart Failure. *Eur Heart J, 22*, 1527-1560.

The Task Force of the Working Group on Heart Failure of the European Society of Cardiology. (1997). The treatment of heart failure. *Eur Heart J, 18*, 736-753.

The Task Force on Heart Failure of the European Society of Cardiology. (1995). Guidelines for the diagnosis of heart failure. *Eur Heart J, 16*, 741-751.

Timmis A D. (1996). Investigation in general practice of patients with suspected heart failure [letter; comment]. *Heart, 75*(6), 642-643.

Vancheri F, Alletto M & Curcio M. (2003). Is clinical diagnosis of heart failure reliable? - Clinical judgement of cardiologists versus internists. *Eur J Intern Med, 14*, 26-31.

Watkins C, Harvey I, Langley C, Faulkner A & Gray S. (1999). General practitioners' use of computers during the consultation. *Br J Gen Pract., 49*(442), 381-383.

Wheeldon N M, MacDonald T M, Flucker C J, McEndrick, A D, McDevitt D G & Struthers A D. (1993). Echocardiography in chronic heart failure in the community. *Q J Med, 86*, 17-23.

Wigton R S. (1988). Use of Linear Models to Analyze Physicians' Decisions. *Med Decis Making, 8*, 241-252.

Wigton R S, Hoellerich V L & Patil K D. (1986). How physicians use clinical information in diagnosing pulmonary embolism. *Med Decis Making, 6*, 2-11.

Wigton R S, Patil K D & Hoellerich V L. (1986). The effect of feedback in learning clinical diagnosis. *J Med Educ, 61*, 816-822.

Wilhelmsen L, Eriksson L, Svärdsudd K & Caidahl K. (1989). Improving the detection and diagnosis of congestive heart failure. *Eur Heart J, 10*(suppl C), 13-18.

Wolf B. (1999). Vicarious functioning as a central process-characteristic of human behavior. *Brunswik Society web-site: http://brunswik.org/notes/essay4.html*(May).

Wolf B. (2000). Processes of constructing judgments and actions by competent individuals with respect to object orientation: Programmatic ideas in the tradition of Brunswikian thoughts. *Brunswik Society web-site: http://brunswik.org/notes/essay7.html*(august).

Zile M R. (2003). Diastolic heart failure. Diagnosis, prognosis, treatment. *Minerva Cardioangiol, 51*, 131-142.

Zile M R & Brutsaert D L. (2002). New Concepts in Diastolic Dysfunction and Diastolic Heart Failure: Part I: Diagnosis, Prognosis, and Measurements of Diastolic Function. *Circulation, 105*, 1387-1393.

SUMMARY IN SWEDISH – SAMMANFATTNING PÅ SVENSKA

Hjärtsvikt är ett vanligt tillstånd med hög dödlighet och sjuklighet, och ökande prevalens. Allmänläkarna har en viktig roll i omhändertagandet av patienter med hjärtsvikt, både när det gäller diagnostik, behandling, uppföljning och samverkan med specialistvården. Det är svårt att ställa diagnosen kronisk hjärtsvikt, fr.a. i tidigt skede av sjukdomen. Syftet med studierna var att undersöka allmänläkarnas diagnostiska bedömning av patienter med misstänkt kronisk hjärtsvikt. Två metoder från kognitiv psykologi användes: klinisk bedömningsanalys (Clinical Judgement Analysis, CJA) i studierna I-IV och tänka högt i studie V. Fallbeskrivningar baserade på autentiska patienter presenterades, antingen i pappersformat (studierna I-III) eller på en datorskärm (studie V). I studie IV diskuterades de teoretiska och praktiska problem som kan uppstå när man vill skapa en uppsättning representativa fallbeskrivningar för CJA-studier, och erfarenheter från studierna I och II användes som exempel.

I studie I bedömde 27 allmänläkare 45 fallbeskrivningar med avseende på sannolikheten för hjärtsvikt. Fem av fallen var dubbletter. Varje allmänläkares diagnostiska strategi definierades som uppsättningen av statistiska regressionsvikter för de variabler som använts för att beskriva patienterna. Läkarna skilde sig kraftigt när det gällde både diagnostiska bedömningar och diagnostiska strategier, men analysen av dubblettfallen visade att de var konsistenta i sina bedömningar. De variabler som betydde mest för bedömningarna var information om lungstas, hjärtstorlek och tidigare förekomst av hjärtinfarkt. När läkarna tillfrågades om vilka variabler de tagit mest hänsyn till i sina bedömningar övervärderade de betydelsen av dyspné.

I studie II bedömde 27 allmänläkare, 22 kardiologer och 21 medicine studerande 40 fallbeskrivningar med avseende på sannolikheten för hjärtsvikt. Fallbeskrivningarna var baserade på välutredda patienter med valida diagnoser och kvaliteten i deltagarnas diagnostiska bedömningar kunde därför studeras. Det var stora skillnader mellan individuella deltagare både när det gällde diagnostiska strategier och kvaliteten på bedömningarna, men skillnader mellan de tre grupperna kunde inte påvisas. De viktigaste variablerna var information om hjärtstorlek och lungstas. Med hjälp av klusteranalys identifierades tre bedömningsstrategier. I den första klustergruppen var hjärtstorlek den dominerande variabeln, i den andra dominerade lungstas, medan vikterna var mer jämnt fördelade på de olika variablerna i den tredje klustergruppen. En tredjedel av deltagarna hörde till den första gruppen, och de hade gjort bättre diagnostiska bedömningar än de övriga. R^2 var i genomsnitt 70% i studie I och II, vilket tyder på att regressionsmodellen kunde predicera bedömningarna bra. I studie III analyserades samma data som i studie II, men med avseende på egenskaperna hos de fallbeskrivningar där deltagarna var mest respektive minst överens när det gällde de diagnostiska bedömningarna. Ett större antal variabler som kunde tala för hjärtsvikt, ökad hjärtstorlek samt förekomst av förmaksflimmer bidrog till att deltagarna blev mer överens när det gällde bedömningarna.

Utgångspunkten för studie IV var att det ofta rekommenderas att man i CJA-studier använder sig av fallbeskrivningar som är representativa för de bedömningar som görs. I artikeln diskuteras begreppet representativitet och vilka krav det medför när det gäller val av relevant population, relevanta variabler, antal fallbeskrivningar och hur fallen ska presenteras. Två faktorer visade sig vara mest problematiska: Den ofta otillräckliga informationen i patientjournalerna och det önskvärda i att hålla nere antalet fallbeskrivningar. Dessa två faktorer ledde till kompromisser bl.a. när det gällde val av variabler och antal variabler.

I studie V bedömde 15 allmänläkare sex av de fallbeskrivningar som använts i studie II. Data analyserades med avseende på hur olika slags information om fallen användes i de diagnostiska bedömningarna. Information om ekokardiografi (som ej presenterats som variabel i de tidigare studierna) var den information som oftast användes som diagnostiskt argument, men i en tredjedel av bedömningssituationerna användes denna information inte alls. Information om hjärtvolym och lungstas användes också ofta. Information om andra relevanta sjukdomar användes också ofta i läkarnas diagnostiska resonemang, men detta avspeglas inte i guidelines.

De två metoder som använts för att studera diagnostiska bedömningar och diagnostiska resonemang som använts i denna avhandling utgör båda användbara metoder för att studera kliniskt beslutsfattande. Ett möjligt tillämpningsområde är studiet av medicinska experter och studenter, som kan ge kunskap av värde för undervisning. Ett annat tillämpningsområde är utveckling och test av olika beslutsstöd integrerade i datorjournaler, eller utveckling av guidelines.