OUTCOME IN PSYCHIATRIC OUTPATIENT SERVICES

RELIABILITY, VALIDITY AND OUTCOME BASED ON ROUTINE ASSESSMENTS WITH THE GAF SCALE

by

Per Söderberg & Stefan Tungström

Department of Psychology
Umeå University, Umeå, Sweden
2007
ABSTRACT


The general aim of the studies presented in this thesis is to investigate the possibility of using clinical data to measure outcomes in psychiatric outpatient services. The specific aims are to investigate whether routine clinical assessments and ratings are reliable and have adequate validity, and then to use these data to calculate treatment outcomes and explore factors that affect these outcomes.

The main result shows that ratings of global mental health made by clinicians in routine clinical work can be used to evaluate treatment outcomes in outpatient settings. The clinicians responsible for diagnosing and assessing patients used the GAF scale with satisfactory reliability ($ICC_{1,1} = 0.81$) and fair interrater reliability (overall kappa = 0.53) when categorizing main diagnostic groups of the DSM-IV axis I. The GAF scale can thus be used to assess global mental health and to monitor outcomes in clinical settings. However, a GAF culture bias was observed. This bias can probably be corrected with feedback and training.

Psychiatric treatment in outpatient settings had a generally positive effect on patients’ global mental health (ES = 0.65). The overall result when clinical significance methodology was used showed that 28.1% of the patients had recovered and a further 6.6% showed reliable improvement. Patients being treated with psychotherapeutically influenced methods showed a considerably better effect (ES = 1.00). There is a dose of sessions effect that is particularly marked for short treatment episodes. Thirteen sessions are required for 50% of the patients to show reliable improvement. The strongest influence on treatment outcome was whether the termination of a patient’s treatment was planned or unplanned.

In conclusion: Clinical databases can be used to study the outcome of psychiatric services provided they a) include a large number of subjects representing an intention-to-treat perspective; b) the instruments used are clinically relevant and reliable; c) the raters contributing to the data base are motivated to decrease attrition; d) the database includes extensive data
to allow for control of confounding factors; and e) data are collected at critical occasions in treatment, such as at the start of treatment and at discharge from treatment, making it possible to focus on effects. Psychiatric outpatient treatment has a positive effect, but considerable improvements may be possible with more stringent use of psychotherapeutic methods, sufficient doses of sessions, and planned terminations. However, the progress of treatment is also affected by such factors as pre-treatment severity and diagnoses.

**Key words:** reliability, validity, outcome, effectiveness, clinical significance, GAF scale, DSM-IV, outpatients, psychiatry, dose response, services
ACKNOWLEDGEMENT

We want to acknowledge all those who have contributed in different ways to the completion of this thesis.

First, we would like to thank Professor Bengt-Åke Armelius, our supervisor, for his guidance and friendly support in our work. Besides being an excellent scientist and a sensible supervisor, he is a brilliant and humorous person to collaborate with.

Special gratitude goes to the staff of the Department of Psychology at Umeå University from whom we have learned much in our repeated visits to the department over the years.

We are grateful for the financial support of the Center for Clinical Research in Dalarna which made this study feasible.

We appreciated the general support of and many valuable discussions with the managers of the psychiatry department in Dalarna, with special thanks to Bo Sjödin, Göran Källberg, and Lena Mallon.

In particular, we would like to thank our former manager Sören Fogde, who allowed us to start our research project, and our present manager Karin Stikå-Mjöberg, who gave us the time and understanding we needed to complete the study.

We would also thank all the clinicians who were willing to help us by making ratings and assessments in our reliability studies.

Special thanks goes to Dag Gjesteby and Bo Ivarsson, managers in the Psychiatric Development Group of NYSAM, for their inspiring support and for letting us use their database.

We would also like to express special personal thanks to our colleagues at the Psychiatric Research and Development Department in Säter, and especially to Kerstin Hedtorp and Ann-Chatrin Dahlberg for their encouragement, and to Rick Bennison for his practical support.

Finally, this research would not have been possible without the support and generosity of our families. I, Per, would like to thank my wife Lena for her unconditional and genuine support, and our children, Tove, Kalle, Anton, Freja, and Gunnar, for their patience. I, Stefan, would like to dedicate special thanks to Karin Lisspers for her encouragement and support, and of course to our daughters, Julia, Sara, and Kerstin.

Dalarna, January, 2007
Per Söderberg & Stefan Tungström
LIST OF PAPERS

This work actually represents two doctoral dissertations. Since our papers are closely related and constitute a logical unit, we obtained permission to publish our work together. Per Söderberg is responsible for papers II, IV, and VI and Stefan Tungström is responsible for papers I, III, and V. In the following text, the studies will be referred to by their roman numerals.


Papers I and II have been reproduced with the permission of *Psychiatric Services*, American Psychiatric Association.
## CONTENTS

**ABSTRACT** .................................................................................................................. III

**ACKNOWLEDGEMENT** .............................................................................................. V

**LIST OF PAPERS** ........................................................................................................ VI

**CONTENTS** ................................................................................................................... VII

**ABBREVIATIONS** .......................................................................................................... X

**INTRODUCTION** .......................................................................................................... 1

Quality assurance for psychiatric outpatient settings.................................................. 2

**The DSM System** ...................................................................................................... 4
- Axis I ............................................................................................................................. 4
- Axis II .......................................................................................................................... 4
- Axis III .......................................................................................................................... 5
- Axis IV .......................................................................................................................... 5
- Axis V .......................................................................................................................... 5
- Reliability of axis I ....................................................................................................... 6
- Reliability of the GAF scale ......................................................................................... 7
- Validity of the GAF Scale ............................................................................................ 10

**Outcome research** ..................................................................................................... 12
- Efficacy and effectiveness .......................................................................................... 14
- Outcomes measured by the GAF scale .................................................................... 17

**Factors affecting outcome** ....................................................................................... 20
- Common and specific therapeutic factors .................................................................. 20
- Evidence-based treatments: Treatment manuals ...................................................... 21
- Dose of treatment ....................................................................................................... 22

**AIMS OF THE THESIS** ............................................................................................. 24
METHODS ................................................................................................................. 25

Subjects .......................................................................................................................... 25
  The BASP Database .................................................................................................... 25
  The Reliability Database .......................................................................................... 26
  GAF Project Database .............................................................................................. 27

Procedures and Statistics .............................................................................................. 27
  Intraclass Correlation (ICC) ...................................................................................... 29
  Generalized kappa ..................................................................................................... 29
  Effect size and meta-analytic statistics ...................................................................... 30
  Clinical significance ................................................................................................. 32

RESULTS OF THE STUDIES ...................................................................................... 35

Study I .................................................................................................................................. 35
  Purpose ......................................................................................................................... 35
  Results ............................................................................................................................. 35
  Conclusions .................................................................................................................... 36

Study II ............................................................................................................................. 36
  Purpose ......................................................................................................................... 36
  Results ............................................................................................................................. 37
  Conclusions .................................................................................................................... 37

Study III ........................................................................................................................... 37
  Purpose ......................................................................................................................... 38
  Results ............................................................................................................................. 38
  Conclusions .................................................................................................................... 38

Study IV ........................................................................................................................... 39
  Purpose ......................................................................................................................... 39
  Results ............................................................................................................................. 40
  Conclusions .................................................................................................................... 40

Study V ............................................................................................................................. 40
  Purpose ......................................................................................................................... 40
  Results ............................................................................................................................. 41
  Conclusion ..................................................................................................................... 41

Study VI ............................................................................................................................. 42
  Purpose ......................................................................................................................... 42
  Results ............................................................................................................................. 42
  Conclusions .................................................................................................................... 43

DISCUSSION .................................................................................................................... 44
Validity of the GAF .................................................................................................. 44
Reliability of the GAF .............................................................................................. 45
Reliability of DSM axis I .......................................................................................... 48
Treatment outcome ................................................................................................... 48
Dose of sessions ........................................................................................................ 51
Treatment termination ............................................................................................. 53
Clinical databases .................................................................................................... 55
  Intention-to-treat .................................................................................................... 56
  Large databases ...................................................................................................... 56
  Many variables ...................................................................................................... 56
  Reliable and valid instruments ............................................................................. 57
  Different perspectives on outcome ....................................................................... 57
  Internal validity ..................................................................................................... 57
  Motivation and low burden .................................................................................. 57
  Measurement interval ............................................................................................ 58
  Database construction .......................................................................................... 58
  Standard operating procedures .......................................................................... 58
Limitations ................................................................................................................ 59
Conclusions ............................................................................................................. 60
REFERENCES ........................................................................................................... 61
# ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>BASP</td>
<td>BasPsykiatrins databas (Psychiatric outpatient database)</td>
</tr>
<tr>
<td>BPRS</td>
<td>Brief Psychiatric Rating Scale</td>
</tr>
<tr>
<td>CS</td>
<td>Clinical significance</td>
</tr>
<tr>
<td>ES</td>
<td>Effect size</td>
</tr>
<tr>
<td>DSM-IV-TR</td>
<td><em>Diagnostic and Statistical Manual of Mental Disorders</em>, 4th edition, text revision</td>
</tr>
<tr>
<td>GAF</td>
<td>Global Assessment of Functioning</td>
</tr>
<tr>
<td>GAF-CB</td>
<td>GAF culture bias</td>
</tr>
<tr>
<td>GAS</td>
<td>Global Assessment Scale</td>
</tr>
<tr>
<td>GHQ12</td>
<td>General Health Questionnaire 12</td>
</tr>
<tr>
<td>HSRS</td>
<td>Health–Sickness Rating Scale</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass correlation</td>
</tr>
<tr>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>ISO 9000</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ITT</td>
<td>Intention to treat</td>
</tr>
<tr>
<td>LUPP</td>
<td>Lokal Uppföljning/Utvärdering av Praktisk Psykiatri. (Local follow-up in routine psychiatric work)</td>
</tr>
<tr>
<td>MSER</td>
<td>Mental Status Examination Record</td>
</tr>
<tr>
<td>PSAS</td>
<td>Psychiatric Symptom Assessment Scale</td>
</tr>
<tr>
<td>RCI</td>
<td>Reliable change index</td>
</tr>
<tr>
<td>RCT</td>
<td>Randomized clinical trial</td>
</tr>
<tr>
<td>SBU</td>
<td>Swedish Council on Technology Assessment in Health Care</td>
</tr>
<tr>
<td>SCID</td>
<td>Structured Clinical Interview for -DSM-IV-TR</td>
</tr>
<tr>
<td>SCL-90</td>
<td>Symptom Check List</td>
</tr>
<tr>
<td>SOFAS</td>
<td>Social and Occupational Functional Assessment Scale</td>
</tr>
<tr>
<td>SIQ</td>
<td>Swedish Institute for Quality</td>
</tr>
<tr>
<td>SOSFS</td>
<td>Socialstyrelsens författningssamling (General recommendations of the National Board of Health)</td>
</tr>
<tr>
<td>SPRI</td>
<td>Swedish Institute for Health Services Development</td>
</tr>
</tbody>
</table>
INTRODUCTION

General Background

In 2005 the Swedish National Institute of Public Health conducted a general survey of the health of the Swedish population using the general health questionnaire GHQ12 (Boström, 2006). Of the sample of 64,000 persons from 16 to 84 years old, 40,000 (63%) of the sample responded. Some 21% of the women and 15% of the men reported reduced mental well-being. We also know that each year about 3000 persons, 2000 women (1.8% of the adult population) and 1000 men (0.9% of the adult population), apply for psychiatric outpatient treatment in the county of Dalarna, and that there are about 90,000 visits to these treatment centers each year (Nysam, 2005). The question that arises is, Does this psychiatric treatment have a positive effect on patients’ mental health?

Many research studies have focused on this question, but many of them have been clinical trials, or have taken other forms that make the results difficult to generalize to psychiatric clinical practice. Westen and Morrison (2001) observed that the rigorous selection criteria for clinical trials exclude around two-thirds of eligible clients from study. Thompson-Brenner and Westen (2005) also point out that the common exclusion criteria for bulimia nervosa in clinical trials exclude approximately 40% of the patients in a naturalistic sample. In a recent paper by Westen (2006) he showed that the situation was even worse for studies of depression, panic disorders and GAD, where about 70% of the patients who were intended to be included in RCT studies were excluded. Consequently, it seems that it is unwise to generalize from the results of clinical trials to ordinary psychiatric services, making it difficult to answer the question of whether treatment in clinical practice has positive effects. It is understandable that clinicians develop different ways of following-up treatment.

The psychiatric outpatient clinics in Dalarna developed a system in which patients were routinely assessed using the DSM system (American Psychiatric Association, 2000). At the same time, other information about demographic variables and treatment characteristics was recorded. These clinical data were saved in databases to be used for quality assurance and to study the outcome of routine clinical work. In this study we examined 10 years of this collected clinical data. The instruments used to measure outcomes were the DSM system axes I and V, known as the Global Assessment of Functioning Scale (GAF).
Quality assurance for psychiatric outpatient settings

The Swedish Institute for Health Services Development (SPRI) published a report that described how to measure and develop quality in the health care sector (SPRI rapport 230, 1987). This publication was an important step in introducing quality assurance systems into the Swedish health care system. The Swedish Federation of County Councils supported this quality assurance work by introducing a quality assurance system similar to the one developed by the Swedish Institute for Quality (SIQ).

In 1996 a new regulation in the national Health and Medical Service Act laid down that the health care system must systematically and continually ensure the quality of health care. Simultaneously, the National Board of Health published general recommendations on quality systems in health care (SOSFS 1996:24). This document, which was influenced by the international quality system ISO 9000, described how to develop a quality assurance system.

At the end of the 1980s and beginning of the 1990s, there was great interest in developing local quality assurance systems in Swedish psychiatric organizations. In 1988 the SPRI started a program of local follow-up studies and evaluation of routine practice in psychiatry (SPRI-rapport, 1988). In the county of Dalarna, the clinics in Hedemora and Falun where the authors of this thesis worked were actively involved in developing systems for evaluation and follow up. In 1993 we persuaded the commission to develop a wide-ranging quality assurance system for the entire psychiatric outpatient care system in Dalarna (Söderberg & Tungström, 1994).

The system was based on the ideas of structure, process, and outcome developed by Donabedian (1980). Its primary focus was on systematic and ongoing follow-up of the effects of outpatient treatment episodes. It required clinicians at the two pilot clinics to registered their diagnostic assessment of patients after their first and last visits. The diagnosis was to be done in terms of the DSM III-R (American Psychiatric Association, 1987). Global mental health was also assessed with Luborsky’s (1962) Health–Sickness Rating Scale. All visits were recorded using codes assigned to different treatments. At termination, the clinicians made a second assessment of each patient’s functioning using the HSRS and also rated their change on three dimensions (symptom reduction, improvement in social functioning, and improvement in occupational functioning). The difference between the ratings at the first and the last visit was used to estimate treatment outcome. Finally, the type of ending (planned or unplanned) was recorded.
A pilot project in the two clinics ran for two years. After evaluating this project, the county board of psychiatry decided to extend the quality assurances system to the whole county in 1996. The variables were updated and the diagnostic system changed to the new DSM-IV Manual (American Psychiatric Association, 1994). At the same time, the HSRS scale was replaced with the GAF scale. A simple model of the implementation is shown in figure 1.

**Figure 1.** A model of the implementation of the quality assurance system BASP

The house in the model represents the psychiatric outpatient services that each patient receives during treatment. However, this metaphor does not cover all the factors that may influence a patient’s mental health. Since psychiatric treatment can continue for a considerable time, the influence of other factors must also be taken into account. Many patients are also supported by social service agencies, general practitioners, colleagues at work, and, of course, by their families, relatives and friends.

To monitor the quality of psychiatric work, a database was developed that could be used for local follow-up. Different clinics could enter their own reports monthly or half-yearly. The board received an annual quality report (Kvalitetsbokslut) for the whole county (Söderberg 1997, 1998, 1999, 2000, 2001, 2002, Tungström 2003, 2004). These reports described the patient diagnostic groups, the treatment process, and the outcomes.
The follow-up system was named BASP. A more extensive description of it has been published by Söderberg and Tungström (2002).

The DSM System

The Diagnostic and Statistical Manual of Mental Disorders (DSM), was first published by the American Psychiatric Association in 1952 (DSM I). Since then several other editions have been published, with the latest, DSM IV, released in 1994. A subsequent revision which was released in 2000 is known as the DSM IV-TR (Text Revision). The system is used worldwide in clinical work and is intended to help clinicians to plan treatment, predict outcomes and evaluate treatment. It categorizes symptoms and aspects of functioning in different classes of mental disorders. Given that mental disorders are complex, there is no assumption that the classes are discrete, and therefore several diagnoses are allowed. A complete DSM diagnosis require assessment on several dimensions. The present system is constructed as a multi-axial system with five different axes. Each axis refers to a well-defined domain: Axis I – Clinical disorders; Axis II – Personality disorders or mental retardation; Axis III – General medical conditions; Axis IV – Psychosocial and environmental problems; Axis V – Global assessment of functioning.

Axis I

Axis I covers clinical disorders, the diagnoses that are usually associated with psychiatric nosology and other conditions that may be the focus of clinical attention, such as v-codes. If several diagnoses are reported in an inpatient setting, the principal diagnosis is listed as the condition responsible for the admission and is reported first on axis I. In outpatient settings, the reason for a visit is reported as the first diagnosis.

Axis II

Axis II covers personality disorders or mental retardations. The DSM IV defines a personality disorder as a pattern of inner experience and behavior that deviates markedly from the expectations in the culture. The DSM IV categorizes ten different personality disorders, divided into three
different clusters. These conditions are reported on a separate axis to decrease the risk of their being overlooked, for the presence of such conditions is usually significant in clinical work.

**Axis III**

Axis III covers general medical conditions that are relevant to understanding a patient’s psychiatric status. It is important to note any known physical problem and the source of the information about that problem.

**Axis IV**

Axis IV covers psychosocial and environmental problems reported by the patient that may affect the diagnosis, treatment and prognosis of the mental health disorder. The focus is usually on problems or situations involving negative life events in the last year, but even earlier situations can be listed if they affect the patient’s present mental health. The various psychosocial and environmental problems are grouped in nine different categories, several of which can be listed for one patient.

**Axis V**

The Global Assessment of Functioning (GAF) scale was introduced as the new rating scale for Axis V in the *DSM III-R* (1987) and today it is listed as Axis V in the *DSM IV-TR* (2000). This scale measures the global severity of a patient’s psychiatric illness at a particular point in time and facilitates the development of procedures for measuring outcomes. This scale enjoys widespread acceptance in clinical settings and is frequently used by researchers. It may well be the most common and widely used rating scale for assessing the global functioning of patients with psychiatric disorders (Bodlund, Kullgren, Ekselius, Lindström, and von Knorring, 1994; Hall, 1995; Piersma, and Boes, 1997; Moos, Nichol, and Moos, 2002).

The GAF scale was developed from the similar GAS scale (Endicott, Spitzer, Fleiss, and Cohen, 1976), and was influenced by the Health–Sickness Rating scale (Luborsky, 1962).
The GAF scale is a single-item standard instrument designed to measure the global severity of psychiatric illness on the basis of psychological, social, and occupational functioning. The first version in the *DSM III-R* (1987) was a 90-point scale, which was changed to 100-point scale in the *DSM IV* (1994). The change added 10 points in the upper portion of the scale. The scale is divided into 10 anchor intervals, with descriptions and examples of symptoms and of social and occupational functioning. Assessments are performed by giving a patient a score of between 1 and 100, where anchor point 1 represents the most severely ill psychiatric patient and anchor point 100, at the opposite end of the continuum, represent the healthiest person. The assessment of psychiatric illness should not include functional impairments due to physical or environmental limitations.

**Reliability of axis I**

Since diagnostic assessments are essential in treatment planning, it is important that the assessments be accurate. The question thus arises of how frequently clinicians examining the same patient will provide the same diagnosis. Before the introduction of the DSM III system, some reliability studies were performed for axis I. In a field trial by Spitzer, Forman, and Nee (1979), 274 clinicians evaluated 281 patients. More than half of the patients (n = 150) were interviewed and assessed jointly and the rest were assessed in separate interviews. Of the 15 diagnostic groups to which patients were assigned, 7 had kappa coefficients over 0.75 when the interviews were conducted jointly. However, the kappa coefficients for only 4 of the groups were above 0.75 when interviews were conducted separately. With separate interviews, the kappa coefficients for such common diagnostic groups as psychotic disorders, schizoaffective disorders, mood disorders and anxiety disorders was below 0.75. These results indicate that there is usually some disagreement between clinicians. Mental disorders are complex, and there are various options when making diagnostic decisions. There is thus a need for routines that contribute to good evaluations in order to ensure reliability in clinical practice. A structured interview, SCID, has been developed, but this interview is complex and time-consuming, which means that it is unsuited to routine clinical work (First, Spitzer, Gibbon, and Williams, 1997).
Reliability of the GAF scale

Considering the prevalence of the GAF scale in clinical and research work, surprisingly few studies have investigated its reliability. The reliability study referred to in DSM III-R (1987), DSM IV (1994) and DSM IV-TR (2000) uses results obtained from five substudies of the GAS scale (Endicott et al., 1976). They found that the intraclass correlations for the GAS scale varied between 0.61 and 0.91 and that the value of the standard error varied between 5.0 and 8.0 points. Table 3 summarizes studies focusing on the reliability of the GAF scale and shows the number of raters, number of patients or cases, types of patients, and the reliability estimates.
Table 1. Studies of the reliability of the GAF scale

<table>
<thead>
<tr>
<th>Study</th>
<th>Raters</th>
<th>Clinical material</th>
<th>Patient group</th>
<th>Reliability estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hall (1995)</td>
<td>12 Varying occupations, untrained</td>
<td>16 Patient chart summaries</td>
<td>Depression patients in inpatient care</td>
<td>0.62 (intake) 0.90 (discharge)</td>
</tr>
<tr>
<td>Jones, Thornicroft, Coffey &amp; Dunn (1995)</td>
<td>12 Varying occupations, one training session</td>
<td>103 Patients</td>
<td>Mixed group</td>
<td>0.76</td>
</tr>
<tr>
<td>Loevdahl &amp; Friis (1996)</td>
<td>104 Varying occupations</td>
<td>5 Case vignettes</td>
<td>Patients in inpatient and outpatient care</td>
<td>0.54</td>
</tr>
<tr>
<td>Michels et al. (1996)</td>
<td>45 Psychiatrists, psychologists</td>
<td>12 Case studies</td>
<td>Mixed group of patients GAF 10–63</td>
<td>0.65</td>
</tr>
<tr>
<td>Tracy, Adler, Rotrosen, Edson &amp; Lavori (1997)</td>
<td>10 Systematic training, education, research assistants</td>
<td>23 Case studies</td>
<td>Psychotic patients</td>
<td>0.90</td>
</tr>
<tr>
<td>Hilsenroth et al. (2000)</td>
<td>10 Trained raters, psychologists</td>
<td>44 Patients</td>
<td>Outpatient care</td>
<td>0.86</td>
</tr>
<tr>
<td>Oliver, Cooray, Tyrer, &amp; Cicchetti (2003)</td>
<td>25 Varying occupations 2 hours training</td>
<td>38 Case vignettes</td>
<td>Clients with learning disability</td>
<td>0.28</td>
</tr>
</tbody>
</table>

This summary provides a divergent picture of the reliability of the GAF scale. Reliability is lower in the studies performed in a clinical setting (Hall, 1995; Jones, Thornicroft, Coffey, and Dunn, 1995; Loevdahl and Friis, 1996; Michels et al., 1996) than in those conducted in a research setting (Tracy, Adler, Rotrosen, Edson, and Lavori, 1997; Hilsenroth et al., 2000). The study by Oliver, Cooray, Tyrer, and Cicchetti (2003) is an
outlier and shows that the GAF is not reliable enough to be used in the routine assessment of learning disability.

The GAF scale has been criticized because of these inconsistencies in validity and reliability. Several authors have suggested that it should either be revised or eliminated from the DSM system. Goldman, Skodol, and Lave (1992) suggest field testing of a modified version of the GAF scale in which the measures of social and occupational functioning are separated from the measures of symptoms and psychological functioning. This involved the development of a new scale, the Social and Occupational Functional Assessment Scale (SOFAS), which also takes physical limitations and mental impairment into account (DSM IV, 1994).

A key criticism is that a one-dimensional score on the GAF scale poorly describes the complex picture of mental illness. Kennedy (2003) thus introduced a similar scale for global assessment of functioning, the K axis, which offers a multidimensional measure. A wide range of important clinical areas are divided into seven subscales, measured in the same way as the GAF scale, from 1 to 100. This total clinical picture is then reduced to a single number in the form of a GAF equivalent score.

Hall (1995) also criticized the original GAF scale, arguing that the poor description of assessment criteria would result in low reliability. Hall accordingly modified the scale by introducing more criteria and more detailed directions for assigning scores. These two scales, the K axis and the modified GAF, have been found to have reasonably good reliability and are in fact good alternatives to the GAF scale. Yet because these scales are more comprehensive, it takes longer to make an assessment using them, and they have not been as widely adopted as the GAF.

Another development of the GAF scale has seen it treated as two independent subscales, GAF Symptom and GAF Disability. This practice is common in psychiatric clinical practice in Sweden (Ivarsson, Erdner, and Malm, 2006) and Norway (Karterud, Pedersen, Loevdahl, and Friis, 1998). The procedure to be followed is described as follows: “An alternative method is to treat the GAF as if it were two scales, one for symptom severity and another for level of functioning. Using the steps above, make one rating for severity and a second for level of functioning. The worst of the two ratings can be used as the GAF” (First, 1995, p. 259). This procedure has been used in research projects by Jones et al. (1995) and by Malm, Ivarsson, Allebeck, and Fallon (2003). The psychometrics of these two subscales of the GAF have been studied by Jones et al. (1995) and by Pedersen, Hagtvet, and Karterud (2007).
Validity of the GAF Scale

The GAF scale is supposed to measure mental health or illness on a hypothetical continuum taking into account psychological, social, and occupational functioning. Psychological functioning is operationalized in terms of various psychiatric symptoms. These symptoms are not considered diagnostic criteria, for the scale is not supposed to be linked to diagnoses. On the other hand, some relation between diagnoses and the GAF scale is to be expected, for the categories describe different diagnoses on axis I, which deals with symptoms as well as functional matters. Coffey, Jones, and Thornicroft (1996) demonstrate the validity of the GAF scale by reporting an association between diagnoses and GAF scores. Lower GAF scores were associated with schizophrenia, major depression, and personality disorders. Table 2 lists several other studies that have examined the relation between symptoms or diagnoses and GAF scores. As can be seen, all except one of the studies show that symptoms or diagnoses explain some 20% of the variation in GAF scores.
Table 2. *Studies examining the relation between GAF scores and diagnoses*

<table>
<thead>
<tr>
<th>Study</th>
<th>Scale</th>
<th>n</th>
<th>Study group</th>
<th>Explained variation in GAF %</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endicott et al. (1976)</td>
<td>GAS</td>
<td>–</td>
<td>Psychiatric inpatients at admission</td>
<td>19%</td>
<td>Mental Status Examination Record (MSER). Endicott, Spitzer &amp; Fleiss (1975)</td>
</tr>
<tr>
<td>Hilsenroth et al. (2000)</td>
<td>GAF</td>
<td>44</td>
<td>Psychiatric outpatients</td>
<td>21%</td>
<td>Symptom Check List (SCL-90)</td>
</tr>
<tr>
<td>Roy-Byrne, Dagadakis, Unutzer, &amp; Ries (1996)</td>
<td>GAF</td>
<td>337</td>
<td>Psychiatric inpatients hospitalized over an eight-month period</td>
<td>19%</td>
<td>Psychiatric Symptom Assessment Scale (PSAS)</td>
</tr>
<tr>
<td>Yamauchi, Ono, Baba, &amp; Ikegami (2001)</td>
<td>GAF</td>
<td>2461</td>
<td>Psychiatric inpatients</td>
<td>18%</td>
<td>Brief Psychiatric Rating Scale (BPRS)</td>
</tr>
<tr>
<td>Garrison (2000)</td>
<td>GAF</td>
<td>549</td>
<td>Private practice patients</td>
<td>21%</td>
<td>18 clinical subcategories</td>
</tr>
<tr>
<td>Robert et al. (1991)</td>
<td>GAF</td>
<td>78</td>
<td>Psychiatric inpatients</td>
<td>64%</td>
<td>DSM-III-R 7 clinical subcategories</td>
</tr>
</tbody>
</table>
In an earlier study, Jones et al. (1995) focused on the validity of the GAF scale by interpreting the association between variation in GAF scores and changes in patients’ support needs and current medication. Bacon, Collins, and Plake (2002) studied the influence of functional factors on GAF scores and concluded that GAF ratings are influenced more by symptoms than by adaptive functioning. Patterson and Myung-Shin (1995) conducted a field trial to focus on the construct validity of a modified version of the GAF. Testing 196 outpatients, they found that six factors accounted for 51.75% of the variance in GAF scores. These factors were access to and ability to use transportation, degree of social support, current living situation, current potential for violence, medical compliance, and number of agency referrals.

**Outcome research**

Published studies usually focus on limited aspects of treatment such as psychotherapy, medical treatment, patients with different diagnoses, or cost-effectiveness. Posternak, Zimmerman, and Solomon (2002) demonstrated how clinical data can be used to measure outcomes in routine clinical practice. They used clinicians’ ratings and a standardized outcome rating instrument to follow up patients with mood disorders treated with standard antidepressant medication. They showed that these ratings could be used to answer clinical questions that were important to practicing clinicians.

In a review of published reports of outcome evaluations of mental health services, Speer and Newman (1996) noted that much more is written about outcome evaluation in mental health services than is actually done. They identified 39 studies evaluating the outcome of various psychosocial interventions in community settings. While several of these studies suffered from methodological problems due to inadequate sample sizes and attrition, Speer and Newman felt that many clinics had made a good beginning in developing a feasible system for outcome evaluation.

Gilbody, House, and Sheldon (2002) systematically reviewed the application of outcome research in mental health services. They found only nine examples of published outcome research based on analyses of large databases of patient information collected from clinicians in routine clinical practice. Moreover the results from several of these studies were of questionable value to both clinicians and patients as confounding variables were not considered.
Gilbody et al. (2002) discuss the strengths and weakness of outcome research, pointing out that results obtained using highly selected samples in random trials are difficult to apply in clinical practice. Data should preferably be generated while offering routine services rather than in artificially constructed trials. If the data is continuously collected over time, it will be possible to provide answers to clinically relevant questions fairly quickly and to back them up with solid statistics. On the other hand, this sort of data is usually collected as part of an administrative process and focuses on only a few variables, rather than on all the variables that may be significant for research. There is also the problem of confounding variables (such as disease severity) and selection bias, both of which may affect outcomes. When data covers only a few variables, it is impossible to control for such influences. In their review, Gilbody et al. found only two studies that had collected a broad range of clinical outcomes and case-mix variables.

As described by Gilbody et al. (2002), outcome research is not the same as quasi-experimental research. In outcome research, the researchers examine data that has already been collected in clinical practice, whereas researchers conducting quasi-experimental studies implement an intervention in one setting and compare the outcomes for patients who have and have not been subjected to the intervention. However, other researchers intertwine the concepts of quasi-experimental research and mental health services research (Donenberg, Lyons, and Howard, 1999). According to these authors, mental health service research examines a broader range of questions and emphasizes large databases and natural settings. The studies are often designed to test the effectiveness of standard interventions in clinical settings that treat a heterogeneous group of individuals with multiple problems. The goal of mental health services research is to improve the quality, impact, and cost-effectiveness of the delivery of mental health services (Newman, Howard, Windle, and Hohmann, 1994). To fulfill this goal, clinical data must be collected in databases that can be used to evaluate treatment outcomes and to support mental health policy. The design of these databases must be influenced by both the traditional research paradigm of clinical trials and the ongoing development of mental health services research.

The problems identified above have meant that few studies have focused on outcomes in mental health services in general. There is, however, a trend toward obtaining more generalizable evidence on treatment effects in order to bridge the gap between traditional research and clinical practice. But the most common approach to research in clinical practice is still either to examine the efficacy of treatments
delivered in RCT studies or to focus on the effectiveness of treatments in studies designed to enable the results to be generalized to clinical practice.

**Efficacy and effectiveness**

When studying the effects of psychiatric treatment, it is important to address the concept of validity. Cook and Campbell (1979, p.37) defined validity as the best available approximation of the truth. The truth is a complex concept and depends on several factors. Thus one can invoke several different types of validity, depending on the setting in which the effects are being studied. Two types identified by Campbell and Stanley (1963) are internal and external validity. Internal validity is present when two variables have a causal relationship, or when the absence of a relationship causes an absence of causality. External validity refers to the extent to which the causal relationship can be generalized to other settings. These different kinds of validity are associated with two different kinds of treatment effect studies: those focusing on efficacy and those focusing on effectiveness. Efficacy studies involve well-controlled trials under conditions that usually differ from normal clinical work. Effectiveness studies are conducted in clinical conditions in which experimental conditions are not available (Kazdin, 1998). The two standpoints should be regarded as being on a continuum, with efficacy studies is at one end and effectiveness studies at the other.

In treatment studies, the kinds of question the research focuses on strongly influence the method and design. When focusing on causal relations (e.g., Is treatment A better than treatment B?) randomized controlled designs are required. However, such a design does not in itself guarantee internal validity. Mulder, Frampton, Joyce, and Porter (2003) list five threats to internal validity. First, small samples increase the risk of Type I and Type II statistical errors. Second, patients are not similar, even if we have strict inclusion criteria. Although randomization of subjects will help to reduce the effect of variations, psychiatric patients differ in a number of ways that may influence the result. Third, psychiatric problems have a high spontaneous recovery rate, making it difficult to argue that recovery is due to a specific treatment. Fourth, different outcome measures and differences in when the outcome is measured will give different results. For instance, the result will probably differ depending on whether it is assessed immediately after the last session or six months after the end of treatment. Fifth, different treatments are not all that different.
Several important factors are common to different therapies, making it difficult to specify exactly what causes the results.

Seligman (1995) argues that standard efficacy studies cannot provide empirical validation of psychotherapy since they omit many crucial elements of what is done in the field. For example, in clinical work therapy continues until a patient shows marked improvement. The termination is usually planned together with the patient. In efficacy studies, however, the design of the study usually specifies a set number of sessions. Patients who do not benefit from the therapy will probably drop out before this number is reached. In clinical work, psychotherapists also usually try several methods in order to determine which method best suits the patient. The process is self-correcting in that if one technique does not work, another is tried. The patient often has an opinion on what method they prefer and believe in. In addition, patients usually have multiple problems, even if the therapist and the patient have agreed to focus only on the one they have identified as the main problem. After the initial assessment, the therapist who will best fit a patient is assigned to a case, or the patient chooses the therapist. But in efficacy studies, the therapist is assigned at random. These methodological issues mean it is difficult to generalize the result from efficacy studies to clinical practice.

There are also some threats to external validity. Since the main purpose of studies that desire external validity is to generalize the findings to clinical conditions, the threats to external validity are related to factors that limit the generalizability of the results. It is important to be very clear about what exactly the sample represents in terms of patients, therapists, settings, context and so forth. External validity is also threatened by non-completers, that is, patients who do not complete treatment. Failure to include this group in a study makes it impossible to generalize the findings to general clinical conditions. Any treatment effect will probably be overestimated and will only apply to those who complete treatment. Consequently Kendall, Flannery-Schroeder, and Ford (1999) suggest that intention-to-treat analyses be used when assessing the effectiveness of treatment. Such analyses produce conservative estimates which take into account the outcomes of all participants. This focus is relevant in clinical practice where all the patients who apply for treatment have to be assessed and treated.

In conclusion, both efficacy and effectiveness studies are needed to improve clinical work: new methods have to be developed by focusing on causal effects; old ones have to be improved, and all methods have to be evaluated when used in general clinical work. The main differences between the two different types of studies are summarized in table 1.
Table 3. *The differences between efficacy and effectiveness studies*

<table>
<thead>
<tr>
<th></th>
<th><strong>Efficacy studies</strong></th>
<th><strong>Effectiveness studies</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research question</strong></td>
<td>Does the treatment work under special conditions?</td>
<td>Does the treatment work under clinical conditions?</td>
</tr>
<tr>
<td><strong>Validity</strong></td>
<td>Internal validity</td>
<td>External validity</td>
</tr>
<tr>
<td><strong>Research methods</strong></td>
<td>Randomized clinical trials</td>
<td>Naturalistic settings, quasi-experimental designs</td>
</tr>
<tr>
<td><strong>Focus of results</strong></td>
<td>Causal relations</td>
<td>Generality</td>
</tr>
<tr>
<td><strong>Type of sample</strong></td>
<td>Small well-defined samples</td>
<td>Intention-to-treat</td>
</tr>
<tr>
<td><strong>Treatment characteristics</strong></td>
<td>Well-defined Manualized</td>
<td>Integrated Self directed (what works)</td>
</tr>
<tr>
<td><strong>Therapist characteristics</strong></td>
<td>High adherence Trained in the particular method</td>
<td>Low adherence Responsive to patient’s needs Eclectic</td>
</tr>
</tbody>
</table>

Lagomasino, Dwight-Johnson, and Simpson (2005) discuss the fact that psychiatric research almost always involves randomized clinical trials with highly selected samples. They argue that evidence from real-world studies promotes high-quality, cost-effective mental health care.

There have been several attempts to combine the advantages of the two perspectives. Wells (1999) points out that efficacy and effectiveness studies in psychiatry rely on different and often competing design strategies and approaches to analysis. He recommends that researchers using different approaches cooperate on integrated hybrid studies with a quasi-experimental design. Carroll and Rounsaville (2003) also suggest a hybrid research model, linking efficacy and effectiveness research in substance abuse treatment. Their model retains important ingredients of efficacy research while also addressing effectiveness issues such as
diversity in settings, less restrictive inclusion criteria, and the satisfaction of patients and clinicians.

Patient-focused psychotherapy research has also developed as an alternative to efficacy and effectiveness research (Lutz, Lowry, Kopta, Einstein, and Howard, 2001; Lutz, 2002, 2003). This type of research is based on large clinical data sets provided by several outpatient clinics or by health insurance companies. Its theoretical origins lie in the dosage and phase models of psychotherapy (Howard, Kopta, Krause, and Orlinsky, 1986; Howard, Lueger, Maling, and Martinovich, 1993), which focus on the positive relationship between the number of sessions and improvement. Sperry, Brill, Howard, and Grissom (1996) and Lambert, Harmon, Slade, Whipple, and Hawkins (2005) have presented methods of monitoring patients’ progress through treatment. These methods make it possible to determine whether a patient is responding to the treatment that is being given. The patient’s condition is assessed during treatment in order to give feedback to the therapist as to whether progress can be expected for this particular patient. This approach makes it possible to test the knowledge derived from efficacy studies in clinical settings.

Outcomes measured by the GAF scale

One purpose for including the GAF scale in the DSM-IV system was to obtain information about individuals’ overall level of functioning. This information is intended to be used to plan treatment, measure its impact, and predict its outcome (DSM-IV TR, 2000). Many mental health service organizations have now adopted this scale as a standard tool. Patients’ GAF ratings are usually saved in databases together with other clinical data. These records offer an opportunity to study the outcome of mental health services. There are many clinical examples of the use of the GAF scale in large-scale evaluations of psychiatric services. In Sweden, the Quality Star project (Ivarsson et al., 2006) developed a tool for generic outcome monitoring with follow-up measurements from different perspectives. In this project, the GAF scale is used as two separate scales: GAF Symptoms and GAF Disability.

Several research studies in different settings with different research designs have reported on psychiatric outcomes measured by the GAF scale. Dufton and Siddique (1992) explored the use of the GAF scale in a day hospital setting. The GAF scale was found to be useful for categorizing patients and detecting change from admission to discharge. A group of 71 patients with mixed diagnoses were assessed for admission to
the day hospital program. Of this group, 12 were excluded from the study for reasons such as illiteracy or refusal to complete questionnaires. The mean length of stay at the day hospital was 99.2 days overall (SD = 71.8), with a range from 14 to 381 days. All the GAF ratings were done by the psychiatrist at the day hospital, who was trained in the use of DSM-III-R. The mean GAF at admission was 49.8 (SD = 6.9), and at discharge, 55.3 (SD = 8.1). The result shows a significant improvement of 5.5 GAF points between admission and discharge (paired sample test, $t = -4.61$, $p = 0.0001$). This study has some limitations, such as the fact that all the ratings were done by the same psychiatrist. On the other hand, this reflects what happens in many clinical settings.

Piersma and Boes (1997) analyzed three psychiatric samples (adult inpatients, adult partial hospitalization patients, and adolescent inpatients) at admission and discharge from psychiatric treatment. All the patients in the samples were rated using the GAF scale in routine clinical practice and all the samples showed significantly less dysfunction at discharge. The mean change between intake GAF score and discharge GAF score was approximately 15 points for adult inpatients, 10 points for adult partial hospitalization patients, and 22 points for the adolescent inpatients. These scores may have been inflated by the fact that the clinicians had access to the admission scores when making discharge ratings.

Garrison (2000) used the GAF scale in a study of psychoanalytic psychotherapy in private practice. The goal of the study was to describe the GAF characteristics of patients and to develop path analytic models for predicting changes in GAF scores. In this effectiveness study, 551 patients were rated with two GAF scores, one representing their functioning at the beginning of treatment and the other at the time of a research interview. The change in GAF score was calculated from these scores. The mean change in GAF score was 16.7 (SD = 12.6) and the median was 15 GAF points. A major weakness in this study was that the assessment of the initial GAF score was made retrospectively.

Malm et al. (2003) evaluated the effectiveness of two community-based programs, namely integrated care (IC) and rational rehabilitation (RR) for patients with schizophrenia and other psychotic disorders ($N = 84$). Seven different measurements contributed to the assessment of outcome. The GAF scale used was divided into GAF Symptom, GAF Disability, and GAF DSM-IV total. The major finding at two-year follow-up was that the IC group showed significantly higher social functioning than the RR group, as measured by GAF Disability scale and WHO-DAS scores. The mean change in GAF Disability between baseline and follow-
up was 7 GAF points for the IC group compared to 1 GAF point for the RR group.

The Department of Veterans Affairs has for several years routinely collected clinical GAF ratings from different psychiatric clinics all over the United States. Greenberg and Rosenheck (2005a, 2005b) use data from the Veterans Health Administration (VHA) to show how the GAF scale can be used as an outcome measure in outpatient and inpatient facilities. Their first study was based on a sample of 283,754 GAF ratings from both inpatient and outpatient clinics. It provides empirical support for cautious use of the GAF scale and suggests that it might be used to assess outcomes at individual facilities over time. This result comes from a large database collected by health care systems on an ongoing basis, with low quality control of the data. The raters were untrained clinicians and there was no formal assessment of interrater reliability. The authors suggest that a quality assurance system for GAF scoring with training and testing for reliability should be implemented across facilities.

In their second study, Greenberg and Rosenheck (2005b) used the GAF scale to show positive, statistically significant associations between several continuity-of-care measures and clinical outcomes. As in the first study, they used three large samples of GAF ratings (n = 8,334, n = 49,946, n = 123,371) from the Department of Veterans Affairs database. The result shows small, statistically significant differences in outcome, although some effects may not be clinically meaningful.

In a recent study, Philips, Wennberg, Werbart, and Schubert (2006) evaluated the outcomes for young adults in psychoanalytic psychotherapy in a naturalistic setting. They used various outcome measures, including the GAF scale, at the start and termination of treatment. The cohort consisted of 134 patients, aged between 18 and 25 years old with various problems, who applied for psychotherapy and agreed to participate in the research study. The various measurements agreed in showing significant positive change between the start of therapy and its termination. The largest improvement was identified by the GAF ratings, which showed a change from a mean intake score of 57.3 to a mean treatment termination score of 67.3, giving an effect size of 1.50.

Moos and his colleagues (2002) studied patients with psychiatric or substance abuse disorders and found that GAF ratings were only minimally associated with treatment outcomes. They found no robust association between GAF ratings and outcomes assessed by clinicians in an interview or by patients’ self-report at follow-up. Their result shows that patients’ diagnoses and symptoms were stronger predictors of their GAF ratings than their social or occupational functioning. They thus
questioned the value of including GAF ratings in a program for predicting the allocation and outcomes of mental health care.

To summarize, the GAF scale is widely used all over the world both in routine clinical practice and in efficacy studies. It is comprehensive and easy to use, and can be reliable if the raters have been educated in the use of the scale and use it regularly. It is an instrument that can be used to evaluate psychiatric care in routine clinical practice.

Factors affecting outcome

In outcome research it is important to ask what causes the improvement. It is unlikely that just one condition or one specific process determines whether treatment is effective. There are probably many independent and interacting variables. We usually talk about moderated and mediated effects when a third variable is involved.

A moderating variable is one that influences the direction or the strength of the relationship between an independent and a dependent variable (Baron and Kenny, 1986; Holmbeck, 1997). For example, if a specific treatment is shown to be more effective for women than for men, gender is the moderator variable between treatment and outcome. In other words, there is an interaction between outcome and gender.

A mediator is a variable that specifies how a given effect occurs, a process or mechanism through which the independent variable is able to influence outcome. One example of a mediator variable that been proven to influence outcome is the therapeutic alliance, the collaborative bond between the therapist and the patient. For example, Meyer et al. (2002) show that expectancies (patients’ expectation of the effectiveness of treatment) and outcome are mediated by the patients’ contribution to the alliance.

Common and specific therapeutic factors

Research on psychotherapy usually distinguishes between common and specific factors that cause improvement. The common factors are not specific to any particular technique and include the expectation of improvement and a therapeutic relationship involving trust, warmth, understanding, acceptance, and kindness (Asay and Lambert, 1999). These common factors have also been labeled as non-specific factors or
placebos. The unique factors are defined as the specific features of a particular technique which make treatment effective.

Lambert (1992) developed a graphic model of the factors that influence the outcome of psychotherapy. The model also indicates the degree to which different factors contribute to outcome. He identified four main factors from the research literature: client variables, the therapeutic relationship, expectancy, and psychotherapeutic techniques. Client factors such as a client’s ego strength, motivation, and the severity of disturbance account for 40% of the improvement in psychotherapy. The therapeutic relationship, which includes factors such as empathy, positive regard, non-possessive warmth, and acceptance, which are found in a variety of therapies regardless of the theoretical orientation, account for a further 30% of the improvement. The third factor is the expectation of good results, which explains 15% in the model. In this study Lambert treated expectancy as equivalent to a placebo effect. Finally, the last factor, psychotherapeutic techniques, accounts for 15% of the improvement.

Research does not provide strong evidence for the specific effects of psychotherapy. Wampold (2001) identified four different research designs that have been used to examine specific effects: component designs, comparative designs with placebo controls, designs that examine mediating effects, and designs that examine the interaction between person and treatment. He summarized the research conducted using these designs and concluded that it offered no evidence for the existence of specific effects and no indication that one ingredient was better than another.

Evidence-based treatments: Treatment manuals

To improve treatment in clinical practice and implement methods supported by research results, summaries of research results have been produced for clinicians and decision makers. These reports present treatment methods that are backed up by research. In 1993 the Cochrane Collaboration (www.cochrane.org) was established to disseminate systematic reviews about the effect of healthcare interventions based on evidence from clinical trials and other studies. In Sweden the Swedish Council on Technology Assessment in Health Care (www.sbu.se) conducts scientific assessments to identify which interventions offer the greatest benefit to patients. The Campbell Collaboration (www.campbellcollaboration.org) has also been established to disseminate
systematic reviews of studies in the social, behavioral, and educational fields.

There has also been a trend to create manuals explicating different therapeutic ingredients or treatment techniques. These manuals describe how a particular theoretical orientation sees a specific disorder and outline specific ingredients of the treatment method. This trend toward producing treatment manuals for specific diagnostic groups and state-of-the-art documents for empirically supported treatments derives from a medical model. Within the framework of the DSM system, additional reviews have been published with clinical guidelines for establishing diagnoses. A case book published by Spitzer, Gibbon, Skodol, Williams, and First (2002) includes evidence-based suggestions about the treatment methods to be used for each patient described. The American Psychiatric Association (2000) has published guidelines developed by psychiatrists who are actively engaged in clinical practice. They offer practical guidelines for psychiatric evaluation and psychiatric care for ten diagnostic groups.

Several handbooks and guidelines have also been published to help practicing psychotherapists. One example of this type of book is the *Clinical handbook of psychological disorders*, edited by Barlow (2001). This handbook describes the best-known psychological treatment protocols for the fourteen most frequently treated disorders. Readers can follow the treatment process session-by-session in a clinical application and study a description of theoretical and empirical knowledge produced by clinical researchers and clinicians. Roth and Fonagy (1996) examine research evidence in order to build a scientific base for the clinical practice of psychotherapy. Their review of psychotherapy research sets out evidence relating to the treatment of a wide range of psychiatric conditions and draws conclusions and sets out guidelines for the delivery of mental health services to specific diagnostic groups. In Sweden, the Swedish Council on Technology Assessment in Health Care has published several volumes of guidelines for specific diagnostic groups, including guidelines for mood disorders (SBU-rapport, 2004) and anxiety syndromes (SBU-rapport, 2005).

**Dose of treatment**

How much treatment is enough? Or in other words, how many outpatient sessions are sufficient to achieve a satisfactory result in clinical practice? This question is basic to treatment planning. It has been investigated for a
long time, and early findings shows a positive link between the outcome of psychotherapy and the length of the therapy. In a comprehensive review of the literature, Orlinsky, Grawe, and Parks (1994) found 156 published papers on this topic, the majority of which (100 papers) showed a positive association between length and outcome.

Howard et al. (1986) introduced a new technique for studying the effect of treatment dosage and introduced the concept of the “dose-effect relationship”. They defined the “dose” of sessions of psychotherapy as the unit of treatment in the same way as the dose of medicine is the unit of treatment in pharmacological treatments. Howard et al.’s meta-analytic study covered 30 years of psychotherapy research including over 2,000 patients and showed that the therapeutic effect of psychotherapy is a function of the number of sessions. They described the relation between dose and effect as a negatively accelerating curve with rapid initial gains followed by a slow increase with a higher dose of sessions. On the basis of pre- and post-treatment scores, they showed an overall dose effect such that 53% of patients improved by session eight.

Other studies have repeated these results with different measurements and settings. Kopta, Howard, Lowry, and Beutler (1994) studied a sample of 854 patients who were in psychotherapy at five health centers and found that 11 sessions were required before 50% of the patients improved. Kadera, Lambert, and Andrews (1996) showed that 50% of patients had improved after 16 sessions. Hansen and Lambert (2003) suggest that 15 to 18 sessions are needed for the same level of recovery.

The standard number of sessions for psychotherapy treatment in naturalistic settings is rather low. Hansen, Lambert and Forman (2002) reported mean and median scores from six different naturalistic settings (N = 6,072) and showed an overall low mean (4.3) and median (3.0) numbers of sessions per treatment episode. A literature review by Garfield (1994, table 6.2, p.193) showed that the median number of sessions reported by studies from 1940 to 1989 was between 4 and 10. These findings about the dose-effect relationship highlight the importance of the question of whether psychiatric patients in outpatient settings are receiving sufficient treatment.
AIMS OF THE THESIS

The general aim of the present series of studies was to investigate the possibility of measuring psychiatric outpatient outcomes using clinical data. The specific aims were to investigate whether clinical routine assessments and ratings were reliable and have adequate validity, and then to use this data to calculate treatment outcomes and determine whether certain factors influence the treatment effect. The objectives of the individual studies that made up this project were as follows:

- To investigate the validity of the Global Assessment of Functioning scale when routinely used in clinical work (study I).
- To investigate the reliability of the Global Assessment of Functioning scale and analyze factors that affect measurement errors when the scale is used by regular psychiatric staff (studies II and IV).
- To investigate interrater reliability for the main diagnostic groups of the DSM-IV axis I when routinely used in psychiatric clinical practice (study III).
- To investigate the outcome of ordinary outpatient psychiatric treatment (study V).
- To examine whether there is any dose of sessions effect in psychiatric outpatient’s treatment in routine clinical service (study VI).
METHODS

Subjects

The data in the different studies were collected from three different data sources: the BASP database used for follow-up and quality assurance in public psychiatric outpatient care in the county of Dalarna, Sweden; the reliability database of psychiatric outpatient raters in the county of Dalarna; and the GAF Project database compiled as part of a national psychiatric project in which 40 outpatient clinics in Sweden participated. Table 4 summarizes the descriptive data for the participating subjects in the different studies and the time-period over which the data were collected.

Table 4. Description of sources of data

<table>
<thead>
<tr>
<th>Study</th>
<th>Database</th>
<th>Sample</th>
<th>Time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>10,234 patients</td>
<td>5,538 patients, first visit ratings</td>
<td>1997 – June 2000</td>
</tr>
<tr>
<td></td>
<td>BASP database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>81 raters</td>
<td>81 raters, different occupations</td>
<td>March 2002</td>
</tr>
<tr>
<td></td>
<td>Reliability database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>81 raters</td>
<td>70 raters, different occupations</td>
<td>March 2002</td>
</tr>
<tr>
<td></td>
<td>Reliability database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>31,416 patients</td>
<td>28,554 patients, single visit ratings</td>
<td>1999–2003</td>
</tr>
<tr>
<td></td>
<td>GAF project database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>19,353 patients</td>
<td>11,966 patients, pre- &amp; post ratings</td>
<td>1994 – June 2003</td>
</tr>
<tr>
<td></td>
<td>BASP database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>19,353 patients</td>
<td>11,966 patients, pre- &amp; post ratings</td>
<td>1994 – June 2003</td>
</tr>
<tr>
<td></td>
<td>BASP database</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The BASP Database

In 1994 a database to continuously collect clinical data about the treatment of psychiatric outpatients was established in the county of
Dalarna. It was intended to be used for treatment evaluation, outcome studies, and quality assurance. All patients were registered and assessed on all five axes of DSM-IV at the start of treatment and on axes I, II and V of DSM-IV at treatment termination. Clinicians also recorded the dominant method of treatment at every session with each patient. Seventeen sites, all part of the public mental health system in the county of Dalarna, contributed to the database. The county has 275,000 inhabitants and all the sites had the same mission – offering all outpatient psychiatric treatment in their district.

In study I, a sample of 5,538 patients was drawn from the 10,234 patients entered in the BASP database between 1 January 1997 and June 2000. We then excluded 1,795 of these patients because they had no GAF rating and 1,027 because the record lacked a primary diagnosis. Six sites with a total of 1,875 patients were also excluded from the sample because they had fewer than five patients in one or several diagnostic groups.

In studies V and VI, an intention-to-treat sample of 11,966 treatment episodes with both pre- and post-treatment assessments and ratings was drawn from the BASP database. The entire database contained 19,353 episodes, but 7,059 patients were excluded from the study because they were still in treatment in June 2003. A further 193 patients who had no reported visit and 135 who died during treatment were also excluded.

The Reliability Database

The sample of raters in the reliability studies (studies II and III) was drawn from the population of 144 clinicians who were all the available raters who made the psychiatric assessments recorded in the BASP database during March 2002 when the study was conducted. A total of 81 raters participated in the GAF reliability study (study II) and 70 of these 81 raters participated in the DSM axis I reliability study (study III). The difference in sample size was caused by internal attrition. The participating sites represented all geographic sectors of the psychiatric care system in the province. Some sites chose not to participate for various reasons, including management not giving priority to the studies or staff not having time to participate.

To control for the attrition of raters, the population of the 144 raters was examined. It was found that the distribution of occupations did not differ between the population and the participating raters. It was also found that the 81 participating clinicians in the GAF reliability study had used GAF for an average of 4.8 years, compared to 4.7 years for all 144
raters. In the DSM axis I reliability study, the distribution of occupations in the sample of raters was the same. However, these raters had longer experience of using the DSM, having used it for an average of 5.6 years compared to 4.7 years in the population. It can be concluded that the samples of raters in both studies represent the population of raters.

**GAF Project Database**

The GAF Project database contains data from a national psychiatric co-operative project, the GAF Project, in which 40 outpatient clinics in Sweden participated. These clinics all had defined catchment areas and were responsible for outpatient psychiatric care in their area. These areas varied from 8,000 to 93,000 inhabitants and were located in small or medium-size towns in the southern or central parts of Sweden. Due to the size of the catchment areas, the prevalence of psychiatric illness can be considered equivalent. The clinics used the GAF to rate all patients during the same two weeks in early March and registered each patient’s clinical data. If a patient was seen more than once during the two-week period, the lowest GAF value was used. The same procedure was followed for five consecutive years and the data were registered in a database to be used for comparison of patient characteristics and GAF values between clinics. A total of 98.5% of all patients who received treatment at the participating clinics during the five-years by two-week period were included in the database. The total database consists of 31,416 patients and the number of patients per clinic ranged from 76 to 526. All patients were diagnosed by DSM-IV criteria (Axis I and Axis II). Patients whose diagnoses were not defined by the major diagnostic categories were excluded (n = 2,862) in study IV, leaving a sample of 28,554 patients.

**Procedures and Statistics**

Several statistical methods were used in the studies including variance analytic statistical methods, r-statistics including linear regression models, and meta-analytic statistics. These methods are related and are suitable for use with large samples. Table 5 lists the methods and briefly describes what they were used for. All analyses were performed using SPSS versions 10.1.4 to 14.0 (2005).
<table>
<thead>
<tr>
<th>Study</th>
<th>Methods</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>Eta-squared statistics</td>
<td>Association between independent variables and initial GAF ratings</td>
</tr>
<tr>
<td></td>
<td>Linear multiple regression model</td>
<td>Calculate $R^2$ to show the amount of unique variance in GAF ratings explained by the independent variables</td>
</tr>
<tr>
<td></td>
<td>Variance component model, ANOVA type 1</td>
<td>Investigate the interaction between the independent variables</td>
</tr>
<tr>
<td></td>
<td>Eta-squared statistics</td>
<td>Exemplify the association between different diagnostic groups and GAF.</td>
</tr>
<tr>
<td>Study 2</td>
<td>ICC</td>
<td>Investigate the interrater reliability of GAF ratings</td>
</tr>
<tr>
<td></td>
<td>ANOVA three-way analysis of variance</td>
<td>Analyze whether the independent variables (attitude, experience, frequency) were associated with the reliability of the GAF ratings</td>
</tr>
<tr>
<td></td>
<td>One-way ANOVA</td>
<td>Analyze whether occupation is associated with reliability of GAF ratings</td>
</tr>
<tr>
<td>Study 3</td>
<td>Generalized kappa</td>
<td>Investigate interrater reliability on the DSM-IV axis I</td>
</tr>
<tr>
<td>Study 4</td>
<td>Linear regression model</td>
<td>Analyze main diagnostic group as the independent variable and GAF scores as the dependent variable</td>
</tr>
<tr>
<td></td>
<td>Paired sample t-tests</td>
<td>Analyze group difference between mean of GAF-CB (GAF culture bias) before and after feedback as dependent variable and feedback group as independent variable</td>
</tr>
<tr>
<td>Study 5</td>
<td>Effect size</td>
<td>Study the effect of ordinary outpatient psychiatric treatment by measuring pre- and post-treatment GAF scores</td>
</tr>
<tr>
<td></td>
<td>Eta-squared statistics</td>
<td>Analyze how the categorical variables independently influenced the effect size</td>
</tr>
<tr>
<td></td>
<td>Linear regression model</td>
<td>Analyze the relationship between treatment effects and other factors</td>
</tr>
<tr>
<td>Study 6</td>
<td>Clinical significance</td>
<td>Analyze the effect of treatment</td>
</tr>
<tr>
<td></td>
<td>Linear regression analysis</td>
<td>Analyze the dose of sessions effect.</td>
</tr>
</tbody>
</table>
Intraclass Correlation (ICC)

The most important factors affecting the result of a reliability study of clinical raters are the chosen design, the statistical calculations performed, and the sample of patients and raters included. Many methods of calculating reliability are based on variance analytical models that focus on sources of variance. In studies where different raters assess the same cases, calculation of interrater agreement is central. There are several different formulae for calculating intraclass correlations (ICC). A classic paper by Shrout and Fleiss (1979) showed how different formulae yield different reliability assessments, with a variation of 0.17 to 0.71 for the same data.

In statistical calculations of interrater agreement, the sources of variance are the intrarater variance (the variation in the raters’ assessments of the same patient) and the interpatient variance (the variation between different patients). Reliability is generally calculated as the quotient between these sources of variance: the interpatient variance divided by the total variance (interpatient variance and intrarater variance). The relative magnitude of the interpatient and intrarater variance is important because it makes the calculation of intraclass correlations sensitive to the study design. If patients with large differences in the severity of their illness (large interpatient variance) are used in a reliability study, the reliability assessment will have a higher value than if the researcher had chosen patients with a similar global severity of illness profile (small interpatient variance). The situation is similar with respect to the variation among raters (the intrarater variance), but the effect is reversed, with a smaller difference among raters yielding a higher reliability estimate.

In both reliability studies, studies II and III, the patients and the raters included were considered representative of the psychiatric outpatient population of patients and raters.

Generalized kappa

ICC statistics are not usually used to study the reliability on categorical variables such as psychiatric diagnoses. The simplest way to calculate the equivalence with which different clinicians assess patients is to calculate the ratio between the number of similar decisions and the number of possible decisions. However, there is always the possibility that a clinician may have made the right decision by chance; and this possibility has to be
taken into account. The kappa coefficient developed by Cohen (1960) is a
measure of the proportion by which rater’s similar decisions exceed
chance. The kappa formula is written as follows:

\[ k = \frac{p_o - p_c}{1 - p_c} \]  

(1)

where \( p_o \) is the proportion of observed agreement and \( p_c \) is the
proportion of agreement expected by chance. The kappa coefficient varies
between –1 and 1. A negative value represents agreements that are less
than would be expected by chance, zero represent an agreement that is
equal to what would be expected by chance, and a positive value
represents the amount of agreement that is greater than chance, with 1
representing perfect agreement. Fleiss (1981) suggests that kappa values
above 0.75 should be considered as indicating excellent agreement, values
between 0.40–0.75 as indicating fair to good agreement, and below 0.40 as
poor agreement. Kappa statistics are simple to calculate with two raters
and dichotomous data. However, at times it is desirable to calculate kappa
when two or several raters assess polychotomous data. Bartko and
Carpenter (1976) have suggested methods of doing this.

Effect size and meta-analytic statistics

The method of estimating effect size originally came from the meta-
analytic statistics that refer to statistical methods in synthesis (Cooper &
Hedges, 1994). There are a number of meta-analytical statistical methods,
which Rosenthal (1994) divides into two different families. The \( r \) family
includes such statistical methods as Pearson’s correlations (\( r \)), explained
variance (\( r^2 \)) and \( \eta^2 \), while the \( d \) family includes measures of effect-size
such as Cohen’s \( d \) or Hedge’s \( g \). The \( r \) and \( d \) families are, of course, related.
Rosenthal expresses this relationship as follows:

\[ r^2 = \frac{d^2}{d^2 + 4} \]  

(2)

and

\[ d = \frac{2r}{\sqrt{1 - r^2}} \]  

(3)
To calculate the treatment effect size, the formula for Cohen’s $d$ (formula 4) has been used, but has been modified with the pooled standard deviation as recommended by Lipsey and Wilson (2001), (formula 5).

$$d = \frac{\bar{x}_1 - \bar{x}_2}{sd}$$  \hspace{1cm} (4)

$$\text{treatment effect} = \frac{GAF_{\text{post treatment}} - GAF_{\text{pre-treatment}}}{sd_{\text{pooled}}}$$  \hspace{1cm} (5)

There are several ways to interpret effect size. One way is to convert the effect size to $r^2$ so that it can be interpreted as the proportion of variance in the dependent variable that can accounted for by the independent variable.

Another way to interpret the effect size is suggested by several authors (Cohen, 1988; Glass, 1976; Wampold, 2001; Lambert and Ogles, 2004; and Lipsey & Wilson, 2001). Their suggestion is that the overall effect size can be interpreted as how great part of the pre-treatment distribution will be exceeded by the mean of the post-treatment group. In other words, how many of the non-treated patients will the average treated patient exceed at termination. Using a z-distribution, the effect size can be interpreted as the distance between the pre- and post-treatment means in terms of the standard deviation.

A further way to display the practical importance of an effect is the Binomial Effect Size Display (BESD) (Rosenthal and Rubin, 1982). The BESD can be calculated by using the correlation coefficient to calculate the success rate of treated patients as $0.50 + r/2$ and the success rate of untreated patients as $0.50 - r/2$. The correlation coefficient corresponding to a particular effect-size can be calculated by using formula (2) and then taking the square root of $r^2$.

Cohen (1988) suggests different ranges to interpret effect-size: ES $\geq 0.20$ is a small effect, ES $= 0.5$ a medium effect, and ES $\geq 0.80$ a large effect.

The advantage of a standardized ES value is that it enables comparisons across studies but a more straightforward way to interpret the effect size for a patient is to subtract the patient’s pre-treatment GAF.
score from the post-treatment score. If this GAF change score is equivalent to half a standard deviation in the GAF pre-treatment distribution, the effect is medium, and if it is equal to or greater than one standard deviation, the effect is large.

Clinical significance

In the early 1980s, researchers began to the question how to analyze and report the results obtained from research on the outcome of psychotherapy. They were hesitant about evaluating treatment effects solely in terms of statistical significance tests of the mean difference between groups. Such tests inform us that differences between groups are greater than would be expected by chance, but have little to do with the clinical importance of the effect for an individual. Researchers thus started to discuss the results in terms of “social validity” (Kazdin, 1977), that is, whether the treatment effect is of clinical or practical importance, and “clinical significance”, that is, the extent to which clinical outcomes achieve a clinically meaningful change. Kazdin (1980, p. 272) defined clinical significance as “whether the improvement enhances the client's everyday functioning and, essentially, ameliorates the problem for which the client sought treatment”. Jacobson, Follette, and Revenstorf (1984) pointed out that many different definitions of clinical significance had been used in the literature and that the concept was often poorly defined. They proposed two criteria for determining clinical improvement: normative groups (dysfunctional and functional) and statistically reliable improvement.

This two-step method for determining clinically significant change was further improved by Christensen and Mendoza (1986) and summarized and republished by Jacobson and Truax (1991).

To qualify as clinically significant (CS) and reliable change, the following criteria must be met. First, at the start of treatment, the patient must be among the dysfunctional clinical population and must improve sufficiently to pass the cutoff point and enter the functional population. Second, the improvement must be statistically reliable in the sense that it must be larger than the scale’s measurement error. This latter criterion is referred to as the Reliable Change Index (RCI). With these two specific criteria, a patient’s outcome can categorized into one of four groups: recovered (passed both CS normative groups and the RCI criteria), improved (passed the RCI criteria alone in a positive direction), unchanged (passed neither CS nor RCI) or deteriorated (passed RCI in a negative direction).
Jacobson and Truax (1991) used the formula shown in equation (6) to calculate the cutoff point \( c \) between the functional and dysfunctional groups. The cutoff defines the point that statistically divides the two normative populations.

\[
c = \frac{SD_0 M_1 + SD_1 M_o}{SD_0 + SD_1}
\]  

(6)

\( M_0 \) represents the mean of the functional group; \( SD_0 \), the standard deviation of the functional group; \( M_1 \), the mean of the dysfunctional group; and \( SD_1 \) the standard deviation of the dysfunctional group. The RCI can be calculated as the minimum difference between a patient’s pre- and post-treatment score, when we account for measurement error. The formulas used are those of Jacobson and Truax (equations 7, 8, and 9).

\[
S_E = SD \sqrt{1 - r_x}
\]

(7)

\[
S_{diff} = \sqrt{2(S_E)^2}
\]

(8)

\[
RCI = \frac{GAF_{post-treatment} - GAF_{pre-treatment}}{S_{diff}}
\]

(9)

The value of the standard error of the measurement (\( S_E \)) is estimated from the measurements reliable index (\( r_x \)) and the scales standard deviation (SD).

Finally, the confidence interval for the RCI must be chosen. Wise (2004) suggests that certain rating scales may be used with varying confidence levels, for example, 80% or 90% confidence rather than 95% confidence. This approach obviously risks overestimating the results, but it does provide an opportunity to interpret trends that is especially useful in clinical practice. Wise recommend a confidence interval of 80% or 90%
when clinicians are using a measurement for global assessment of functioning, like the GAF scale.

The development of the clinical significance method has decreased the gap between clinical research and clinical practice, since its primary interest is the degree of change for the individual patient. The method has, however, also been criticized by Hsu (1989) and Speer (1992) for not taking account of regression toward the mean and introduced new alternative methods. Hageman and Arrindell (1999) introduced a third method using another equation to calculate reliable change that corrected for regression to the mean. They also pointed out the importance of using the correct value when calculating the standard error of measurement, which can influence the comparability of research findings across studies. Finally, Speer, and Greenbaum (1995) presented a fourth method, hierarchical linear modeling, which is a multiwave data approach that can be used when there are more than two assessments per individual between pretest and posttest. Bauer, Lambert, and Nielsen (2004) compared these methods of calculating clinical significance using a sample of 386 outpatients treated in routine clinical practice. They found that the result differed depending on which statistical method was used to calculate it. The method developed by Jacobson and Truax (1991) had the advantage of producing a moderate estimate of the treatment effect, being easy to compute and having cutoff estimates available for many instruments. Therefore Bauer et al. (2004) recommended that this method be used for outcome studies.
RESULTS OF THE STUDIES

Study I


Purpose

The Global Assessment of Functioning (GAF) scale is used to judge an individual’s overall level of functioning and to measure the outcome of treatment. However, the usefulness of the scale depends on the reliability and validity of the regular clinical assessments.

Before using GAF data from clinical databases to measure treatment outcome, it was important to study the validity of the scale and find out whether it could be used as a comprehensive measure of mental health. We approached the question of validity by using data from a clinical database and focusing on the relationship of the GAF to the other axes of the DSM system. The purpose was to see whether systematic variation in the assessments matched what would be expected on theoretical and empirical grounds.

Our hypotheses were (a) there is a relationship between GAF scores and diagnoses on axes I and II; (b) there will be no relationship between GAF scores and general medical conditions on axis III; (c) there will be a relatively strong relationship between GAF scores and psychosocial problems on axis IV. We also investigated the association between the GAF and sociodemographic variables extrinsic to the DSM system. In addition, we demonstrated that the amount of systematic variance in GAF scores explained by axis I depended on the selection of diagnostic groups in the sample.

Results

This study showed that the variation in the GAF scores matched what would be expected. Seventeen percent of the systematic variance in GAF scores was explained by diagnostic differences as defined on DSM-IV axis I, and 5.1% by psychosocial and environmental problems as measured on
axis IV. Unexpectedly, the site variable explained another 3.6% of the variance. The study also showed that the amount of variance in GAF scores explained by axis I differed substantially, from between 52.5% to 17.0%, depending on the number of diagnostic groups and which groups were included in the sample.

Conclusions

The study showed the expected systematic variation in the scale and is in line with several other studies of the routine use of GAF scores in clinical work over several years. GAF scores are related to several important factors such as diagnoses, symptoms, functioning, and psychosocial problems. The GAF scale can thus be considered a global measurement of mental health. This internal consistency within the axes of the DSM-system supports the validity of the GAF scale.

Study II


Purpose

Patients who are referred to the psychiatric care system are routinely assessed with diagnostic instruments, and their sociodemographic data are also collected. In many clinics, clinicians use the GAF scale to measure outcomes, recording scores as part of the initial interview and at the termination of treatment. It is important to know whether such GAF ratings made by regular clinical staff meet satisfactory standards of reliability.

The purpose of this study was to investigate the reliability of the GAF and analyze certain factors that could produce measurement errors when the scale is used by regular psychiatric staff. A total of eighty-one clinical raters from various psychiatric outpatient clinics in Dalarna, Sweden, rated eight case vignettes using the GAF scale. Interrater reliability was assessed by using intraclass correlation coefficients (ICCs).
Factors associated with reliability were analyzed using raters’ unique residual values.

**Results**

The staff members responsible for assessing first-time patients at outpatient psychiatric clinics appear to be using the GAF with satisfactory reliability (ICC_{1,1} = .81). This result can be generalized to other raters in outpatient psychiatric settings who routinely and continually use the GAF scale. If GAF ratings are used to compare aggregated data, the precision is high enough to detect difference between groups of patients. GAF ratings can thus be used to compare aggregate pre- and post-treatment data for a group of patients or to compare levels of GAF ratings for different diagnostic groups.

The raters’ motivation to use the GAF scale and other measurement instruments in outpatient psychiatry was the strongest factor associated with reliability. If a rater has a positive attitude in this regard, the reliability of GAF ratings is higher. This is an important finding for clinical settings because responsibility for performing GAF assessments is not always associated with any incentive or feedback. Clinical facilities should consider increasing the motivation to perform good assessments.

**Conclusions**

GAF ratings made by clinical psychiatric staff in outpatient settings can be used to measure changes and outcomes in groups. The GAF can be applied as a follow-up instrument in routine clinical settings and used for clinical research or evaluation in systematic quality assurance systems. To increase the reliability of GAF ratings, several raters can be used on each occasion and the average of all ratings can be taken as the final score. It is also important that the ratings be done by motivated clinical raters.

**Study III**

Purpose

The purpose of this study was to investigate whether psychiatric staff could make diagnostic assessments according to DSM IV axis I with acceptable reliability. Given that the DSM diagnoses were assessed and recorded for most of the patients, we wanted to know whether these diagnoses could be used for outcome evaluations. We also wished to investigate whether the ability to use the DSM reliably depends on occupation and experience and whether the reliability differed at different sites. A design with many raters representing the mix of occupations in a multidisciplinary psychiatric outpatient team and a mix of patients that was representative of outpatients was used. A total of 70 raters participated in the rating of 8 cases. The cases were presented either as written case vignettes or as a video of a clinical interview. After with each case was presented, the raters had to choose between several pre-selected diagnostic groups based on the main diagnostic groups in the DSM-IV system. Reliability was calculated using the formulas for generalized kappa with polychotomous data.

Results

The estimate of overall reliability with the generalized kappa proved to be 0.53 when all cases and raters were included. The 95% confidence interval around the overall kappa was estimated to be 0.40–0.67. There were differences in reliability that could be ascribed to occupation or educational level, as well as some differences between sites. No differences due to experience with the DSM-IV system could be observed.

Conclusions

The results show that ordinary staff members at psychiatric outpatient facilities can perform an initial screening according to the diagnostic main groups in DSM-IV axis I with a fair degree of reliability. However, an increase in the reliability is desirable. According to Rosenthal and Rubin (1982), an increased of a mere 0.10 points (e.g. from 0.53 to 0.63) would
increase the proportion of reliable diagnoses from 76.5% to 81.5%. This 5% difference would mean that about 150 more of the approximately 3000 patients who annually apply for treatment at psychiatric outpatient clinics in Dalarna would have reliable diagnoses. Improvements are desirable and can be achieved by clinical routines that assure stringent use of multidisciplinary teams, suitable education, and the use of those occupational groups that are most competent in conducting these assessments.

**Study IV**


**Purpose**

A special type of systematic measurement error, a form of assessment culture bias, has been discovered within groups of raters using the GAF scale. Raters at different clinics tend to systematically rate patients with higher or lower GAF points than would be expected. Such systematic GAF culture biases can cause problems as the ratings are used for a variety of clinical purposes and decisions. Data with this type of systematic measurement bias may lead to wrong conclusions being drawn when patient outcomes are compared across clinics.

The purpose of this study was to analyze this phenomenon and to quantify an estimate of this systematic bias (GAF-CB). A second purpose was to examine whether this bias can be corrected by giving raters at a specific sites feedback on how their ratings compare to the rating profiles of raters at other sites. To investigate this systematic bias, we used a large database of 28,554 GAF ratings of patients from 40 different psychiatric clinics in Sweden. Those GAF ratings were collected during the same time period for five consecutive years. To calculate the size of the systematic bias, we used a regression model and estimated a value for the GAF culture bias (GAF-CB) of each site. The change in the GAF-CB between measurement occasions was measured.
Results

A GAF culture bias does exist, in the sense that raters at some sites systematically rate patients as having higher or lower GAF scores than expected in comparison with other clinics. The magnitude of the bias may decline if raters are given feedback on their results. The result was not complete for all sites, and several clinics still deviated from the expected GAF level after having been given feedback.

Conclusions

GAF raters may develop a systematic site specific bias when the scale is routinely used in clinical practice. Precautions should thus be taken when using the GAF scale to compare different sites. To reduce this bias and ensure reliability, continuous feedback should be delivered to the raters as part of a quality assurance system.

Study V


Purpose

The main objective of this study was to empirically study the outcome of ordinary outpatient psychiatric treatment. Different sources of variance that might explain the outcome were explored to see whether any particular ingredients in clinical work were more influential than others.

A clinical intention-to-treat database containing data from 11,966 patients assessed by clinicians as routine practice in psychiatric outpatient settings in Sweden was used. All the patients were assessed using the GAF scale, both at their first visit and at termination. The GAF ratings were used to calculate the effect size. The influence of psychotherapeutically influenced treatments was compared with that of other methods for different diagnostic main groups. Finally, a hierarchical linear regression model was used to analyze the relationship between the treatment effects
and other factors when variables relating to the patients’ characteristics were controlled.

Results

The main result shows an overall effect-size of 0.65. Several factors were identified as influencing the outcome. Psychotherapeutically oriented treatments were more effective for all diagnostic main groups except psychosis and schizophrenia. The total effect size doubled for those who received psychotherapeutically oriented treatments rather than other treatments at most sessions. Whether the termination of treatment was planned or unplanned also influenced the effect size. Planned termination was the single factor with the strongest positive influence in the study. After controlling for the patients’ characteristics, treatment methods and the method of termination of treatment explained 17.2% of the treatment effect.

Conclusion

Ordinary psychiatric outpatient treatment has a positive effect for most patients. In the study 39% of the patients showed a large effect, 9% a medium to large effect, 12% a medium to small effect, 3% an uncertain effect. However, 36% of the patients showed no effect, and for 1% treatment had a negative effect.

The study showed that the patients’ mean scores at termination were better than the scores of 74.2% of the patients at the start of treatment. The increased effect size for patients receiving psychotherapeutically influenced methods (from the overall 0.65 to 1.00) would statistically increase the number of successfully treated patients by 6.9% if all patients received that kind of treatment. If the termination was planned, the number of successfully treated patients would increase by an additional 2.5%.
Study VI


Purpose

Psychotherapy research has showed a positive association between the number of sessions and the outcome. This association is usually described as a negatively accelerated curve. If this result can be generalized to psychiatric outpatient settings, it can effect an improvement in the quality of routine clinical practice. Thus this study set out to answer the question of whether psychiatric outpatient clinics deliver an optimal dose of sessions to patients in outpatient psychiatry programs.

We investigated whether there was any “dose of sessions effect” in psychiatric outpatients treatment in routine clinical practice. We also examined whether this effect is confounded with other variables such as the severity of the patient’s condition pre-treatment, psychiatric diagnoses, and length of treatment. We used a clinical intention-to-treat data sample, containing data from 11,966 patients assessed by clinicians in psychiatric outpatient settings in the county of Dalarna, Sweden. All patients were assessed using the GAF scale at their first visit and at termination to allow for analysis of clinically significant change.

Results

The main result shows a general dose of sessions effect comparable to that reported in the literature. The effect takes shape as a negatively accelerated curve with significantly better effects up to dose-groups 16 to 20 sessions, and thereafter only marginal improvements in outcome. The dose of sessions curves differ for different diagnostic groups, with patient with more severe diagnoses needing more sessions to gain the same treatment effect as those with lesser disorders. The effect is somewhat confounded with pre-treatment severity.
Conclusions

This study showed that a dose of sessions effect exists and is especially marked in the short treatment episodes in routine clinical work in psychiatric outpatient settings. A large group of patients with short treatment episodes would probably benefit greatly from slightly longer treatment.
DISCUSSION

The aim of this entire study was to investigate the possibility of using clinical follow-up data to measure the outcome of psychiatric outpatient treatment. If this proves to be possible, we can get an answer to our essential question: Does the psychiatric treatment offered to patients have a positive effect on their mental health? This study shows that the assessments routinely made by clinicians using global measurements of mental health can be used to follow up psychiatric work. This finding offers scope for investigating other clinical questions: What treatment works for what kind of patients? What factors influence outcome? What dose of sessions is needed for successful treatment?

Even more important is the fact that the answers suggest ways of improving treatment in the real clinical world. Clinicians can be given feedback that can be used to correct the treatment process, choose treatments that work, and phase out methods that prove inadequate. To make this feedback as effective as possible, it should be given during the treatment process, making it possible to adapt the treatment of every single case. On an organizational level, outcome studies can be used to implement adequate treatment methods and develop routines that support the treatment process. They can inspire improvements in the quality of psychiatric outpatient services. The results from this study show that it is possible to use clinical data for these purposes provided that the measurements are reliable and valid and that there is a clear routine for collecting data.

Validity of the GAF

We began by studying the validity of the GAF scale and found that it can be used as a comprehensive measure of global mental health in routine clinical work. The relationship between GAF scores and the other axes in the DSM IV system was as expected. The relationship to other factors that were not supposed to influence GAF scores was also as expected, except for the site factor. Since the sample was representative of patients in psychiatric outpatient treatment and the ratings were done by ordinary staff, the systematic variation obtained was valid under clinical conditions. The internal consistencies within the DSM axes and the other variables studied supported the validity of the GAF.
However, another question arises: How valid is the GAF scale compared with patients’ opinion of their own mental health? Bodlund et al. (1994) showed that patients could rate themselves satisfactorily using a self-report GAF scale. The correlation between the ratings made by 73 patients and 2 professional raters (clinicians) was $r = 0.62$. The correlation varied from 0.45 to 0.91 depending on the diagnostic group. In general, patients scored themselves lower than the clinicians did. This result shows that clinicians’ and patients’ perspectives vary systematically. There are also some slight differences due to diagnostic group, with patients with adjustment disorders and mood disorders being the most prone to underestimate their scores. To improve the validity of the measurement of patient’s mental health, complementary measures are desirable, especially measurements that patients can use to rate themselves.

In recent years the GAF scale has sometimes been divided into subscales giving a GAF Symptom score and a GAF Disability score. This procedure gives two perspectives on mental health and makes the description more nuanced. Recalculating the data from our study II shows that this procedure is usable, for the reliability of GAF Symptom was calculated as ICC 0.78 and for GAF Disability it was ICC 0.76.

### Reliability of the GAF

The next step was to investigate whether the GAF could be reliable when used by clinicians in routine clinical practice. The result showed satisfactory reliability for aggregated GAF ratings, for example, when comparing the average GAF ratings of groups of patients. The reliability of the ratings was not high enough for them to be used to reliably assess change in individual patients. However, the study also showed that reliability could be improved if specific procedures were adopted. In earlier studies, researchers found that the rater’s education and training are positively related to the reliability of rating scales (Luborsky & Bachrach, 1974; Dworkin et al., 1990). Our own study shows that the rater’s motivation and attitude toward using rating scales also seem to be important. This finding suggests ways to improve the quality of the routines for conducting ratings in clinical practice. Another strategic way of increasing the reliability of GAF ratings would be to apply the same basic statistical principle that is used for ratings in sports like diving and gymnastics, and have all patients rated by more than one clinical rater, with the average of all ratings being recorded as the patient’s GAF score. This practical application of the Spearman-Brown formula results in
greatly improved reliability. If two raters are used, we calculate a predicted reliability of ICC 0.90, and with three raters the ICC would rise to 0.93 rather than the ICC 0.81 obtained in our study. If we translate the reliability index into GAF points, we have an error in a range of plus minus 8 GAF points for two raters or of 6.5 for three raters, rather than an error of 11 GAF points when there is only one rater (confidence interval set to 95%).

Besides these steps to improve the reliability of individual GAF ratings by clinicians, the interpretation of individual GAF ratings can also be improved. One way to do this is to use different confidence levels. Wise (2004) recommends that confidence levels be set to 80% or 90% for clinicians’ ratings, functioning ratings, and assessments of client satisfaction. Using these lower confidence levels increases the risk for miscalculation, but also offers advantages. Information appraised with a lower level of confidence shows a trend or a good approximation of global functioning, which may not be fully correct but which is better than no information at all. This perspective on interpreting GAF scores can be demonstrated in conjunction with different levels of reliability. In figure 2, the ICC values are shown on the x-axis and the GAF error points on the y-axis. The lines show the corresponding GAF error points and ICC levels for three different confidence intervals: 80%, 90%, and 95%.

*Figure 2. The relation between ICC, confidence interval and GAF points*
An ICC of 0.81 as calculated in our study, and a confidence level of 95%, gives a true GAF somewhere in the range of plus or minus 11 GAF points. With the confidence level set to 80%, the marginal risk is plus or minus 5 GAF points. Applying the same principle, if the reliability increases to ICC 0.90, we have a GAF points range of plus or minus 8 with a confidence interval of 95%, and of 3 GAF points with an 80% interval.

Raters are not the only source of differences in reliability. There are also differences between clinics. In an extensive account of our study II, published in Swedish, Söderberg, Tungström, and Armelius (2004) found that the reliability results between clinics ranged from ICC 0.75 to ICC 0.91. In this study we also discovered a systematic measurement error, “the GAF culture bias”, which is investigated in study IV. This same phenomenon was also observed in study I and by Armelius, Gerin, Luborsky, and Alexander (1991) and Loevdahl and Friis (1996). This bias occurs when raters at certain clinics systematically rate patients higher or lower than expected in comparison to other clinics. This type of systematic bias must be noted when the GAF scale is used in clinical practice and especially when these scores are used as data in comparative studies. In our GAF culture study, study IV, we found that this bias could be measured and also showed that it could be reduced by providing feedback on that bias. As this problem has been observed in many clinics, national collaboration between psychiatric sites has begun. Several videos of patient interviews have been produced with the explicit purpose of increasing raters’ diagnostic skills and training them in the use of the GAF scale. Special feedback is provided for the clinics that receive GAF training to improve their ability to use the GAF scale and avoid developing a GAF culture bias.

To summarize, there are many ways to improve the reliability of clinical GAF ratings, including implementing GAF rating educational programs, increasing the motivation of raters to perform good assessments, and following specific routines with more than one rater for each patient. It is also important to continually check for GAF culture bias and calibrate raters by comparing each site’s GAF ratings with those of other comparable sites. It seems that it would be possible for certain clinics to improve their reliability on routine GAF ratings to as high as ICC 0.90–0.95, the same level achieved in the context of a research program (Tracy et al., 1997). According to Nunnally and Bernstein (1994), scores with this level of reliability can be used to follow up treatment effects for individual patients.
Reliability of DSM axis I

Another important variable in routine clinical practice is the reliability of routine assessment of patients’ diagnoses. Staff members in outpatient clinics routinely make a provisional diagnosis after a patient’s first visit. Our reliability study of DSM IV diagnoses on axis I (study III) shows that staff members at outpatient clinics can make diagnostic assessments in terms of the main diagnostic groups with a fair degree of reliability and an overall kappa of 0.53. Even if this result indicates an acceptable degree of reliability, there is still much room for improvement. Increasing the reliability index from 0.53 to 0.63, would increase the proportion of reliable diagnoses from 76.5% to 81.5% according to the calculation suggested by Rosenthal and Rubin (1982). This increase in reliability would carry over into clinical practice and yield about 150 more reliable diagnoses per year in the county of Dalarna. Such an increase is important when planning patients’ treatment, and especially when choosing medical treatments.

The occupation of the raters was found to be an important factor in reliability: clinicians with more education made better assessments than clinicians with lower levels of education. The difference in reliability between occupational groups was rather large, ranging from a kappa value of 0.42 for a mental hospital nurse to 0.62 for the most trained psychotherapists and 0.65 for consulting psychiatrists. This result suggests that we can probably improve reliability fairly easily by routinely having diagnoses made by people in the occupational groups that showed the best reliability in this area. Other practical possibilities for improving reliability include giving members of the less educated groups special education in diagnostics and using joint interviews, which have been shown to increase the kappa by 0.12 (Spitzer et al., 1979).

Treatment outcome

Our investigation of the validity and reliability of DSM IV axes I and V when used in clinical routine showed that diagnostic assessments on axis I have fair reliability and that GAF ratings have adequate reliability for comparison of groups (Nunnally & Bernstein, 1994). We could thus study the outcome of outpatient treatment based on clinical data recorded by clinicians. The overall outcome of outpatient treatment, based on an intention-to-treat sample, gives an effect size (ES) of 0.65. This result is in
the same order of magnitude as many meta-analyses of the effect of psychotherapy. Statistically, an effect size can be interpreted by using overlapping distributions, as suggested by several authors (Cohen, 1988; Glass, 1976; Wampold, 2001; Lambert & Ogles, 2004; and Lipsey & Wilson, 2001). Using this technique, the overall effect size can be interpreted as what proportion of the pre-treatment distribution will be exceeded by the mean of the post-treatment group. In this study, such an interpretation of the effect size of 0.65 shows that the mean patient at termination was better off than 74.2% of the patients at the start of treatment.

Following Cohen’s (1988) guidelines for interpreting ES values, the values for individual patients can be grouped as follows: a negative ES value shows a deterioration; zero shows no treatment effect; greater than zero to 0.20 an uncertain effect; greater than 0.20 and less than 0.5 a small to medium effect; from 0.5 to less than 0.8 a medium to large effect; and 0.8 and greater a large effect. The distribution of the whole patient group that generated a total effect size of 0.65 is grouped as shown in figure 3.

Figure 3. The distribution of effect size for all patients grouped according to Cohen (1988) (n = 11,966)

Figure 3 is based on an intention-to-treat sample. Approximately 20% of the patients in the no-effect group were patients with missing GAF scores, for whom the ES was automatically entered as zero.
As a comparison, a similar pie chart was drawn up including only those patients who received individual psychotherapy. However, in clinical work patients usually receive several types of treatment, with the result that almost no patients received only psychotherapy. Instead, the chart represents the group of patients who received psychotherapy at more than 50% of their treatment sessions. The distribution of the ES for these patients is shown in figure 4. The ES for this group was 1.00. In this group, 8.3% of the patients in the no-effect group were missing GAF ratings, with the result that their scores were automatically entered as zero.

![Pie chart showing the distribution of effect size for patients who received individual psychotherapy at more than 50% of their sessions (n = 3,854)](image)

**Figure 4.** The distribution of effect size for patients who received individual psychotherapy at more than 50% of their sessions (n = 3,854)

Another comparative approach involves creating a benchmark group, as proposed by Lipsey and Wilson (2001). They suggest selecting an alternative group to compare with the pre-test group. The benchmark post-treatment group that was chosen consisted of those patients with planned termination who were assessed as having no diagnosed psychiatric problem at treatment termination (n = 2,824, GAF mean at treatment termination = 75.05, SD = 8.64). When using this group as an alternative post-treatment group, an ES of 1.51 was obtained, showing that an effect almost two and a half times greater than the obtained overall
effect (ES = 0.65) would have been required for all patients at post-treatment to equal this group.

Using the binomial effect size display (BESD) suggested by Rosenthal and Rubin (1982) and comparing the success rate of treated patients for the whole group (ES = 0.65) with that of patients who received individual psychotherapy (ES = 1.00), the success rate increases by 6.9%, from 65.5% to 72.4%. An equivalent improvement in the county of Dalarna would mean that 250 more patients per year would be successfully treated if they received psychotherapy. This is a main effect without considering the influence of other factors such as diagnoses. It is reasonable to assume that the effect of psychotherapy differs between different diagnostic groups. Patients with psychosis or schizophrenia, for example, showed no increased effect when psychotherapeutically influenced methods predominated in their treatment.

Another method of interpreting effects involves assessing the clinical significance of any changes. By this measure, 28.1% of the patients recovered and an additional 6.6% achieved reliably improved change. A totally of 67 patients (0.6%) deteriorated and 64.7% showed no reliable change. These results are comparable with those of other naturalistic studies of clinically significant change. In a study of 4,761 patients from standard outpatient psychotherapy treatment settings in the United States by Hansen and Lambert (2003), 35.2% of the patients recovered or showed reliably improved changed as compared to 34.7% in our study. The main result of the psychiatric outpatient treatment in Dalarna thus seems to be comparable with the results in other standard treatment settings.

**Dose of sessions**

Another important factor that influences the effect of treatment is the dose of sessions, that is, the number of outpatient sessions. Study VI shows the same dose effect in psychiatric outpatient settings as has been found in psychotherapy research: a negatively accelerated trend in relation to the dose of sessions. Groups that received short treatment episodes show greater variations in effect than groups with longer treatment episodes.

As expected, the dose of sessions was found to be confounded with pre-treatment severity. Patient whose pre-treatment GAF scores were very low showed greater improvement. This effect was not simply an artifact of regression to the mean, for the degree of improvement varied
substantially. When patients with severe pre-treatment conditions were divided into four groups, the largest improvement was not found in the most severely affected group. On the other hand, there were no confounding effect between length of treatment and dose of sessions: the correlation between length of treatment and improvement nearly disappeared when controlling for dose-group. This means that the duration of treatment does not have a unique association with improvement in relation to dose of sessions. In other words, treatment outcome is related to the number of sessions, not the length of the treatment. This result indicates that the effect of the dose of sessions cannot be dismissed as simply spontaneous improvement over time.

The dose effect literature usually investigates the number of sessions needed for 50% of the patients to show some improvement. In study VI we found that a dose of 13 sessions were needed before 50% of the patients in the dose-group had improved. This number is comparable to the numbers found in research on dose effects in psychotherapy (Kopta et al., 1994; Kadera et al., 1996; Anderson & Lambert, 2001). It was also found that patients with severe diagnoses need more sessions to improve than patient with lesser disorders. For patients with severe diagnoses, the dose of sessions curve is more linear then curvilinear, with more treatment producing a better outcome.

It is striking that the median number of sessions per treatment episode in our intention-to-treat sample was 5 sessions, and that the mean dose was 12 sessions. Surprisingly, we found that the low dose groups were large, with the single session group being the largest. There is a gap between the standard number of sessions required for effective treatment and what is routinely offered in outpatient settings. Increasing the dose for certain low-dose patients could probably lead to an improvement in the effectiveness of treatment.

On the other hand, many patients received a large dose of sessions. A total of 755 patients (6.3%) in the intention-to-treat sample received more than 40 sessions and accounted for 42.3% of all outpatients’ visits. Of course, this group of patients included the patients with severe disorders that required long-term treatment, but there are also alternative explanations for the large number of sessions which could result in ongoing treatment without any improvement. This study does not address this question, but it forces us to discuss and develop clear and systematic routines that can be used to continually evaluate and monitor the treatment process. This type of ongoing monitoring and feedback about a patient’s improvement and need of treatment has been tested in different settings (Lambert et al., 2005; Lutz et al., 2001; Lutz, 2002).
In a recent review, Lambert et al. (2005) summarized the results of four studies that investigated the effect of giving timely warnings to therapists when patients deviated from the expected pattern of progress during psychotherapy. The results suggest that this monitoring and prediction of treatment failure improved the outcomes for patients who had a negative response to treatment. The authors recommend widespread adoption of this type of feedback system in routine clinical practice.

Lutz et al. (2001) combined a dose-response model for psychotherapy with growth curve modeling to estimate the dose-response for different diagnostic groups from the perspective of well-being, symptoms, and functioning. Their purpose was to use pre-treatment clinical characteristics to predict a patient’s course of improvement. Ongoing improvement could be assessed by tracking each patient’s actual progress against the progress expected on the basis of that patient’s pre-treatment clinical characteristics. Interestingly, the results from Lutz’s research show the same dose-response as the data drawn from our database. Patients with adjustment disorders and mood disorders show similarly steep log-linear dose-response curves. However, there are also systematic differences between our studies, for although the curve has the same profile, Lutz predicts a larger number of improved patients in every dose group.

**Treatment termination**

A high percentage of the patients in the intention-to-treat sample had an unplanned ending to their clinical treatment episode (24.5%). This situation occurs for various reasons when patients and clinicians do not reach any clinical resolution or agree on termination. Patients in these situations also have a significantly lower level of outcome. Statistics from the database show that patients with unplanned endings had a slightly lower GAF mean (60) at the start of treatment than patients with a planned ending (62). However, the difference in outcome was large, with a mean effect size of 0.31 for unplanned termination compared to a mean of 0.91 for patients with planned termination. This result indicates that a large group of patients start outpatient treatment with psychiatric mental health problems and for some reason interrupt the treatment process without resolving these problems.

Previous research has shown that specific factors can be used to predict which patients are likely to drop-out. Swett and Noones (1989)
found that a high level of paranoid ideation, lack of health insurance, and living less than 15 miles from the psychiatric clinic were significantly associated with dropping out of care. Those factors were combined with substance abuse, divorced marital status, and an absence of fee reduction. This result was calculated from data collected on 142 new patients treated in outpatient clinic in the United States. Young, Grusky, Jordan and Belin (2000) show that patients who left treatment were younger, not living at an institution, more likely to be married, less satisfied with their relationships, and had drug or alcohol problems. Their study was conducted in California in a mental health service setting and with a large sample of 554 dropout patients compared to 1,215 patients who stayed in treatment. A Danish study of 131 first-admission psychiatric patients by Tehrani, Krussel, Borg and Munk-Jørgensen (1996) showed that the dropping out was predicted by such characteristics as living alone, unemployment, young age, and a change of treatment service. The most common reasons for interrupting treatment were dissatisfaction with care (44% of the dropouts) and no need for further treatment (20% of the dropouts). In a study from Italy, Percudani, Belloni, Contini and Barbui (2002) showed a significant effect of the patient’s diagnosis: patients with neurotic and personality disorders were more likely to dropout than others. On the other hand, they found no significant socio-demographic differences between patients who dropped out and patients who were discharged by agreement. In another study from Italy, Rossi et al. (2002) identified dropouts as younger and less likely to be married, and found that the previous length of treatment was shorter. This study was conducted in a community-based psychiatric service and analyzed 44 patients rated as having inappropriate termination. Pang, Lum, Ungvari, Wong, and Leung’s (1996) study of outpatient settings in Hong Kong found that the dropout rate was higher for patients who were married and for those with a job. They suggest that their findings may be explained by arguing that marriage provides social stability, thereby reducing the need for psychiatric services, and that employed patients may be less ill.

Research in different countries and settings thus indicates both similarities and differences when it comes to predictors for dropping out. The common factors include young patients being more likely to drop out than elderly patients. Patients with diagnoses such as personality disorders and substance disorders are found to have a higher dropout rate in some research reports, and patients with neurotic problems also seem more prone to drop out. Young et al. (2000) followed up patients who dropped out of psychiatric care and found that on average they had improved, just as had patients who stayed in treatment. Many of the interviewed patients
who dropped out had recovered from an acute problem and did not need further treatment. However, this research has the shortcoming that more than 50% of the dropout patients in the sample could not be located. These patients may have been severely ill and may have done quite poorly.

Even if the dropout patients in our database had a lower outcome score than patients with an agreed termination, we do not know whether they continued to improve after dropping out. Probably, there are groups of patients who drop out and continue to improve, but there are also patients who deteriorate or simply do not improve. The patients in our database who dropped out were roughly five years younger than those who did not. This group included 10% more men than expected, and patients with substance disorders, personality disorders, and anxiety disorders were overrepresented. These differences were all significant and confirm the results from other studies.

Knowing that some groups of patients are more likely to drop than others, we can be more cautious when, for example, young patients attend treatment, try to minimize their dissatisfaction with care, and develop better treatment strategies for specific diagnostic groups.

Clinical databases

An increased interest in understanding clinical work and its outcome combined with developments in computer technology have resulted in a trend toward extended use of databases in psychiatry. This trend opens up new possibilities for research studies with results that can be generalized to clinical work. Continuous, routine data-gathering by clinicians and accessible storage in databases mean that this data can profitably be used for outcome research, provided that certain conditions are fulfilled. Our experience in this study, together with results from others studies (Sederer, Dickey, & Eisen, 1997; Donenberg, Lyons, & Howard, 1999; and Gilbody et al., 2002) enable us to provide the following checklist of issues to be borne in mind when considering using clinical databases in research.

Intention-to-treat

To be able to generalize the results to routine clinical work, the database must contain all patients who were intended to be treated, even those who drop out or for whom the treatment did not proceed as expected. The
focus must be on naturalistic observations and the effect of standard interventions for all patients with single or multiple problems.

Large databases

The database must include a large number of subjects and involve ongoing collection of data over several years in order to have a relatively heterogeneous group of patient from the real world. A large number of subjects will make the analysis more stable and facilitate the creation of contrasting groups, such as diagnostic groups, treatment groups, and sex groups, thus making it possible to use quasi-experimental designs.

Many variables

The database must include with a large range of variables that can be used to eliminate potential confounding influences. For that reason the database has to contain variables that can be used both to test and to control explanations. It must be possible to structure the data descriptively. The types of data to be collected include 1) patient data (gender, age, sociodemographic variables); 2) diagnostic data (DSM diagnosis, symptoms, functioning, severity); 3) treatment data (the relation between patient and clinician, use of psychotherapy, social interventions, medication, dose of sessions); 4) treatment termination (planned, unplanned) and 5) outcome (pre- and post-treatment ratings). Such data enables researchers to calculate the relation between outcomes and various specific factors. The more alternative explanations that can be controlled for, the better. On the other hand, it can be argued that from an ethical perspective the collection of a large amount of data about a patient’s life and treatment represents a threat to the patient’s privacy. It is thus important to choose to record only those variables whose presence in the database can be justified theoretically and empirically.

Reliable and valid instruments

The measurements used have to be valid and reliable when used in clinical conditions. The reliability of these instruments has to be controlled in clinical conditions and repeatedly tested to ensure reliability over time.
Different perspectives on outcome

The outcome measurements should reflect various perspectives so that the outcome can be validated from different perspectives. For example, the professional rating should be complemented with ratings made by the patient.

Internal validity

The variables used must show internal validity. For example, if the clinician has to record the kind of treatment method used, the methods have to be well defined. The nature of any treatment delivered must be expressed in concepts that are well-defined and familiar to clinicians. Internal validity can be enhanced by using manualized treatments, DSM manuals, SCID diagnoses, or other well known standards.

Motivation and low burden

To prevent internal attrition, the database must enjoy the support of management. All clinicians must understand the potential use and value of their ratings and data. Those responsible for the database must demonstrate these possibilities and periodically deliver outcomes in line with the purpose for which the data has been collected. The problem of keeping people motivated to enter data is a common one and must be addressed by showing leadership and strategic thinking. Since the instruments are used continuously over a long period, it is important that they be easy to administer and require little maintenance, or clinicians’ motivation will wane. The GAF scale is a good example of an instrument that meets these criteria.

Measurement interval

Data have to be collected at standardized intervals so that pre–post measurements are possible. To estimate the effect of treatment, at least two ratings are needed and have to be done at set points, such as at the
start of treatment and at termination. If possible, additional ratings during treatment and a certain time after termination is desirable in order to be able to determine whether an effect is maintained, decreases, or increases. To be able to monitor the treatment process, several measurements must be made over the treatment episode.

Database construction

The database must be constructed in such a way that it is possible to assess outcomes and answer desirable clinical questions. It is important to be able to follow the treatment episode of each patient. Data must thus be capable of being sorted in chronological order. For each patient, the ratings at the beginning and end of each treatment episode must be linked to the various sessions and treatment methods used between these ratings.

Standard operating procedures

The procedures for collecting data and administering ratings or measures must be built into the standard operating procedures of the facility. Assessments and the recording of data should not be an extra task but should be incorporated in a clinician’s daily routine. The integration of quality assurance systems with measurement follow-up systems in computerized records presents a challenge.

In recent years, almost all clinics in Sweden have used computerized record-keeping, and thus there are already clinical databases. The advantages of using such databases are that they are large, have continuous data regarding a large number of variables, and represent all patients who were intended to be treated. However they do have the shortcomings that the variables are seldom well-defined, that clinicians don’t use standardized reporting rules, and that few standardized measurements are used. However, in future, these digital databases can be used for mental health services research if the recommendations made in the checklist above are implemented. To do this, researchers must be involved and must collaborate with clinicians and clinical management when creating this digital world.
Limitations

This study has some weaknesses. The first is that our reliability studies involved short case-vignettes and videos rather than actual psychiatric patients. Consequently the information base for the ratings and assessments was not as rich as it usually is in a face-to-face interview. However, it is difficult to tell whether this fact increased or decreased the reliability of our studies. Hyler, Williams and Spitzer (1982) have shown that the overall kappa values were 0.20 points higher when the diagnoses were based on real interviews rather than case summaries.

A second limitation on our reliability studies is that in a research situation, clinicians are probably more concerned to rate and assess patients accurately than they are in routine clinical practice. Moreover the raters had access to the DSM-IV manual, which is probably not the case in all clinical situations. On the other hand, in clinical work diagnostic considerations are often discussed with colleagues and with a whole multi-professional team before deciding on a diagnostic code and current GAF score. These clinical procedures probably contribute to more valid and reliable assessments and ratings.

A third limitation on this study is related to the use of a naturalistic database with routine clinical assessments and ratings when analyzing treatment effects. The database has irregular attrition that affects its internal validity. There are also shortcomings that affect the study's external validity, such as the fact that the number of patient with psychotic syndromes included in the study is not representative of the number of patients with these syndromes in treatment. This problem arises because of the inclusion criteria for the research sample. Most psychotic patients require lengthy treatment periods, and so were still in treatment.

Fourth, in outcome research, the observed effect cannot be presented as caused by the treatment delivered by the clinic. Several other factors could also have played a role, including support from social services, general practitioners, and families. Those influences can be considered non-controlled confounding factors. However, a result of treatment observed over a great many cases with several different confounding factors can still be considered a main effect. The more confounding factors that can be controlled for, the more secure that evidence will be.

The fifth limitation is related to how we measured outcomes. Scores on the GAF scale were rated by clinicians, with the same clinician usually rating both the before and after score. This situation with clinicians as
raters can present an allegiance problem, meaning that the results may be skewed.

Sixth, the post-treatment ratings were made directly after the last treatment session. Thus we do not know whether the treatment effect remains stable over time. A second follow-up after termination of treatment would have been desirable.

**Conclusions**

The following overall conclusions that can be drawn from this study: The GAF scale can be used in outcome studies to measure global mental health in clinical work. The GAF scale has satisfactory reliability when clinicians assess patients in clinical practice. When looking at the reliability of DSM axis I, only fair overall reliability was observed, showing that there is room for improvement here. A certain GAF culture bias was observed, indicating that there can be local rating cultures at clinics. Consequently comparisons between sites must be done with caution.

As regards the outcome of outpatient psychiatric work, the study showed that psychiatric treatment in outpatient clinics does generate a generally positive outcome for the majority of patients. A dose of sessions effect on treatment was identified, which indicates a significantly better outcome as the number of sessions increases up to 16 to 20 sessions, followed by only marginal improvements with more sessions. The dose of sessions effect differed for different diagnostic groups, with patients with more severe diagnoses needing more sessions for improvement than patients in general.

To summarize: Clinical databases can be used to study the outcome of psychiatric services if the data, measurement procedures, and database meet critical criteria. Psychiatric outpatient treatment has a positive effect, but considerable improvements may be achieved with more stringent use of psychotherapeutic methods, adequate doses of sessions, and planned terminations. However, the progress of treatment differs due to factors such as patients’ pre-treatment severity and diagnoses.
REFERENCES


