Established in 1871
Swiss Medical Weekly
Formerly: Schweizerische Medizinische Wochenschrift
The European Journal of Medical Sciences

Original article | Published 4 March 2011, doi:10.4414/smw.2011.13173
Cite this as: Swiss Med Wkly. 2011;141:w13173

Quality assessment of a randomly selected sample of Swiss medical expertises

A pilot study

Susanna Stöhr[a], Yvonne Bollag[a], Holger Auerbach[b], Klaus Eichler[b], Daniel Imhof[b], Thomas Fabbro[c], Niklaus Gyr[a]

[a] asim, Academy of Swiss Insurance Medicine, University Hospital Basel, Switzerland
[b] Winterthur Institute of Health Economics, Zürich University of Applied Sciences, Winterthur, Switzerland
[c] Clinical Trial Unit, University Hospital Basel, Switzerland

Correspondence:
Susanna Stöhr MD
asim, Academy of Swiss Insurance Medicine
University Hospital Basel
Petersgraben 4
CH-4031 Basel
sstoehr@hin.ch

Summary

BACKGROUND: Considerable criticism has lately been raised by the media regarding the quality of Swiss medical expertises. The present investigation was therefore undertaken to assess the professional quality of Swiss medical expertises. The study was part of a market analysis of medical expertises (MGS study).
METHODS: A sample of 97 anonymised expertises randomly chosen from a total of 3165, collected in the MGS study over a period of 3 months, were evaluated by an international board of medical experts and reviewers, using a stepwise developed questionnaire. Each expertise was independently evaluated by two experts. Data were then tested for plausibility (obvious errors and misunderstandings). The main outcome was the overall quality rating of the expertise that was graded from 1 (very poor) to 6 (excellent) in analogy to the Swiss school grading system. For analysis and interpretation the grades were divided into sufficient (grades >= 4) and insufficient (grades <4).
RESULTS: Overall 19.6% (95% confidence interval: 13.1%; 28.3%) of the expertises were rated to be of insufficient quality. The quality was inversely related to the number of involved medical disciplines, the time relapsed since injury and positively related to the difficulty of the expertise. In addition, expertises in the French and Italian languages were rated superior to those in German.
CONCLUSION: Our results confirm recent criticisms that the professional quality of expertises does not suffice. This is hardly acceptable in face of the financial and personal consequences. There is an obvious need for further research using larger samples and for educational programmes on all levels.

Key words: *quality assurance; disability evaluation; reproducibility of results; legal liability; causality*

Introduction

In Switzerland medical expertises are commissioned either to medical specialists practicing in Switzerland or to specialised institutions such as MEDAS (Medizinische Abklärungsstation) or asim (Academy of Swiss Insurance Medicine) by both social insurances (e.g. disability insurance, SUVA) as well as private insurance companies (accident insurance, loss of income protection insurance or liability insurance).
Besides the formal requirements the validity of expertises depends on their professional quality and the comprehensibility of the conclusions for clients and other interested parties such as lawyers and courts.
It is true that the Swiss Federal Court for Insurances (Eidgenössisches Versicherungsgericht, EVG) has defined the minimal requirements for medical expertises (decision of the Federal Court, BGE 125 V 351, EVG, 6.11.1999) [1], but the content related quality requirements remain less well defined. According to a recently published pilot study of the SUVA [2, 3] on the mainly formal quality of accident insurance expertises and the ensuing reactions in the press, it would seem that expertises show a considerable lack of quality. Considering the increasing number of assessments based on these expertises and the consequences for both insured persons and insurers, there is an urgent need for analysis and improvement. Surprisingly, our literature search (Medline/PubMed) has resulted in a very limited number of studies concerning the quality per se or the assessment of the quality of expertises, either in German or in the Anglo-Saxon speaking area. In particular, it was not possible to identify a validated instrument for testing the quality of expertises (questionnaire) by which the expertises could be systematically evaluated. A search analysis in the archives

of the "Schweizerische Ärztezeitung" (SAeZ) and in the juridical literature produced only guidelines for rheumatology, neurology and psychiatry [4–6], as well as juridical requirements for expertises [12]. More guidelines could be located in specialized books [7–9] and institute's publications [10, 11]. However to date no systematic investigation on the professional quality of medical expertises assessed by an independent review team using a validated test instrument could be found.

The aim of the present extended pilot study was to determine the professional quality in a randomized sample of Swiss medical expertises as judged by independent reviewers and by the clients. Furthermore it aimed at developing a suitable instrument for the evaluation of expertises in practice and for the application in future quality control and research projects.

The present investigation focuses on the overall quality and the evaluation procedure. A detailed analysis of the questionnaire will be dealt with in a separate publication.

The project is part of the comprehensive MGS investigation into the current situation of expertises in Switzerland [13; "Medizinische Gutachtensituation in der Schweiz, Studie zur Einschätzung der Marktsituation und zur Schaffung von Markttransparenz und Qualitätssicherung"]. It was suggested by the Swiss Insurance Medicine Association SIM and was implemented by the Winterthur Institute of Health Economics, Zürich University of Applied Sciences, Winterthur and the Academy of Swiss Insurance Medicine asim, University of Basel.

Study outline and methods

Selection of expertises

Medical expertises for quality assessment were randomly chosen from 3165 consecutive expertises received by Swiss insurances from the time period February 1st until April 30th 2008 and registered for the MGS study in an online setting at one location. The expertises concerned disability insurance, accident insurance, loss of income protection and liability insurance. Random sequence was computer generated and concealed (i.e. the operator was blinded for the patient problem and the name of the medical expert, who had performed the expertise). As recommended by the expert board, we stratified randomisation for insurance area to ascertain assessment of a balanced number of expertises from less frequent insurance areas in the source population. For example, the disability insurance contributed for 77% of expertises in the source population, while liability insurances accounted for only 1%. Thus, we aimed at a ratio of 4 (for accident insurances; to compensate for the low number of liability expertises that cover a similar content): 2.5 (for disability insurance): 2.5 (for loss of income insurances): 1 (for liability insurances). Of 104 randomly chosen expertises, seven charts could not be retrieved by insurances. Thus, 97 expertises were included for quality assessment.

The MGS-study was approved by the ethical committee of the cantons of Basel, Switzerland.

Questionnaires (data sheets)

As no validated assessment tool was available from the literature, a questionnaire for the evaluation was elaborated and agreed upon by the international expert team of 14 persons with specialist knowledge in the field of expertises. The team consisted of twelve Swiss and two German experts equally acting as reviewers. The questionnaire (reference website) was designed to assess the formal aspects (section I) as well as aspects regarding the content (sections II and III), further the degree of difficulty (section IV) and the final overall grading (section V) and included specific questions for different areas of expertise.

For example, in the section concerning the formal characteristics a question was, whether a history had been taken. In the sections concerning the content, it was investigated how the data such as symptoms and signs were collected and how the diagnosis was made and derived. Furthermore assessment of the comprehensibility regarding the diagnoses and the consequences thereof was performed.

The grading system depended on the sections: In section I (yes/no questions) the answer "yes" indicated sufficient quality, "no" an insufficient grading. In the sections II-III questions were graded as good, sufficient or insufficient. Furthermore the reviewers had the additional opportunity in all above sections to mark the answers "?" (judgement impossible) or "0" (irrelevant; of no importance) with a cross . The assessment of the degree of difficulty – section IV – was done according to the ratings A-E (A meaning simple, E meaning very difficult, extremely complex) based on the Swiss Tarmed classification (pricing of medical services) In analogy to the Swiss school grading system, the overall quality (section V) was graded 1 to 6 (6 being excellent, 3 insufficient and 1 very poor). For easier interpretation and analysis, the grades were divided into sufficient (grades 4–6) and insufficient (grades 1–3).

Review process and assessment

The expertises were always assigned to two reviewers by the medical project team of the asim, taking the specialty fields of the reviewers into consideration. For that purpose the expert team was completed by seven additional reviewers in order to increase linguistic competence as well as capacity to cover frequent specialties such as psychiatry and rheumatology/orthopaedics and additional disciplines e.g. vascular and hand surgery. Details of the review team are given in table 1.

All reviewers received written instructions on how to use the questionnaire.

Expertises concerning one or two disciplines (monodisciplinary or bidisciplinary) were given to two specialists with knowledge in the involved fields. Expertises concerning more than two disciplines (polydisciplinary) were assigned at random to two specialists whose specialities were involved in the expertise. By always assigning to two reviewers, it became possible to investigate the congruence of the assessment of quality and the answers to the questions in the questionnaire. The two reviews (the two filled-in data sheets) were transferred into a common data sheet by the asim team and registered in an access database as raw data (*procedure 1*). In this way possible discrepant assessments (discrepancies) of the two reviewers could be revealed. The

following situations were considered as discrepancies: different answers in relation to yes/no questions, more than one grade difference or the transition from sufficient to insufficient grades in the parts of the questionnaire assessing the content and the overall quality.

In the next step, the plausibility check (*procedure 2*), all discrepancies between reviewers were checked regarding plausibility, i.e. explicability (obvious errors, misunderstandings etc). The plausibility check was performed following clearly defined rules. They were approved and written down by the expert team.

To exclude mistakes made by both reviewers in common, answers where both reviewers agreed were randomly checked as well.

Besides the overall quality assessment several other factors with a potential influence on the quality of the expertise were analysed, such as the duration for the processing of the expertise and the duration between the beginning of the injury or the disability until commission of the expertise, the type of expertise (mono-, bi-, polydisciplinary), the area of expertise (type of insurance involved), the grade of difficulty and the language.

Comparison with the quality estimation done by the client

In the online questionnaire of the MGS study the satisfaction of the client with the execution of the expertises was specifically asked for by means of criteria about e.g. the client's "satisfaction with the final conclusions". "Satisfaction with the final conclusions" meant for the clients the argumentation and immanent comprehensibility were rated as "substantiated" and "not well substantiated". In the expertises selected for this study the quality as assessed by the expert team was compared with the quality attested by the clients.

Data analysis

The primary outcome was the binary rating of the overall quality assessment into the categories "sufficient" and "insufficient". All estimation and testing was based on generalized-estimation-equation (GEE) for logistic regression models. Thus it was possible to correctly account for both expert ratings for each expertise. For each predictor an univariate model was estimated and its overall significance was inspected with a score test at a two-sided significance level of 5%. The overall proportion of expertises rated as insufficient was analysed with the same method but with an intercept term only. For categorical predictors all estimates and the corresponding 95% confidence intervals were transformed into proportions of expertises that were rated as insufficient for easier interpretation. All calculations were done using R [14] and the library "geepack" [15].

Results

The present report focuses on the overall quality assessment (section V of questionnaire) and the factors influencing the quality outcome. As previously indicated a detailed analysis of the questionnaire will be dealt with in a separate report.

Characteristics of expertises

Our sample of 97 expertises comprised 38 expertises for accident insurances, 25 for the disability insurance, nine for liability insurances (including five medical liability cases)

Table 1: List of experts and reviewers with their relevant speciality regarding review function.

Experts and reviewers				
Expert/Reviewer	Speciality	Year of first diploma	Years of experience in medical expertises	Preferred language
1	Rheumatology	1985	23	french/german
2	Orthopedics and traumatology	1998	16	german/italian
3	Internal Medicine	1990	17	*Was only expert, has not reviewed*
4	Psychiatry and psychotherapy	1977	39	german
5	Rheumatology	1993	16	*Was only expert, has not reviewed*
6	Psychiatry and psychotherapy	1995	17	german
7	Rheumatology	1994	17	german
8	Psychiatry and psychotherapy	1995	18	german
9	Psychiatry and psychotherapy	1996	nd	*Was only expert, has not reviewed*
10	Orthopedics and traumatology	2008	6	italian/french/german
11	Neurology	1957	55	italian/german
12	Surgery	1975	nd	*Was only expert, has not reviewed*
13	Neurology	1980	30	german
14	Rheumatology	1989	12	german

Additional reviewers				
Reviewer	Speciality	Year of first diploma	Years of experience in medical expertises	Preferred language
1	Psychiatry and psychotherapy	1983	33	french/german
2	Psychiatry and psychotherapy	1992	16	Italian/german
3	Cardiology	1983	29	german
4	Angiology	1986	31	german
5	Psychiatry and psychotherapy	1975	34	german/italian
6	Psychiatry and psychotherapy	1999	13	french/german
7	Hand surgery	1986	12	german

and 25 for loss of income insurances (fig. 1a). Due to the stratified selection process, as described in the methods section, the distribution of insurance areas in our sample (e.g. 24% disability expertises, 36% accident insurance expertises) is different to the source population. Roughly, the MGS study contained 77% expertises for the disability insurance, 10% for accident insurances, 10% for loss of income insurances and about 1% for liability insurances.

The information for the expertise was based on patient contact in 87 cases (89.7%) and on records in 10 cases (10.3%). The MGS study includes a lower number of expertises based on records (only 2.3%); therefore the expertises based on records are somewhat overrepresented in our study.

Figures 1a to 1d show the distribution of the evaluated expertises according to area of insurance (disability insurance, accident insurance, loss of income insurance and liability insurance), the number of involved medical specialities (monodisciplinary, bidisciplinary and polydisciplinary, i.e. type of expertise), the specialities and the language.

The greatest number (N = 73) of expertises were monodisciplinary; 12 were bi- and 12 polydisciplinary. The most common specialist fields – independently from mono-, bi- or polydisciplinarity – were psychiatry, orthopaedics, rheumatology and neurology. Sixty five of the 97 expertises were in the German language, 20 in French, and 12 in Italian. In all, it can be concluded that our sample gives a good representation of the MGS study population (table 2).

Figure 1 a-d: Description of the selected expertises

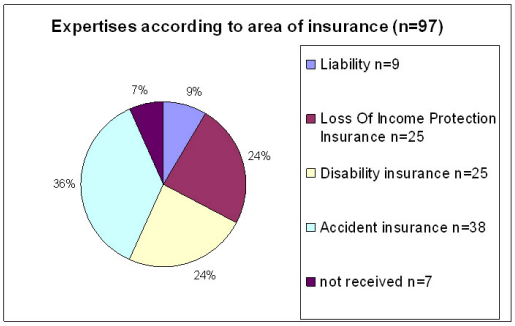

Figure 1a

Expertises according to area of insurance (N = 97).

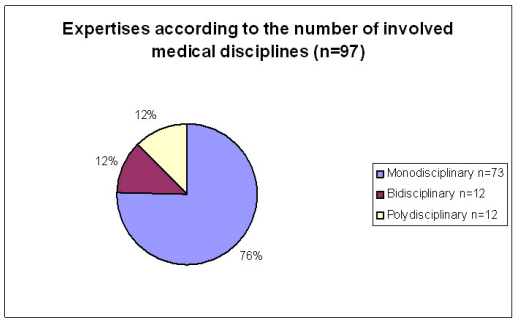

Figure 1b

Type of expertises according to the number of involved medical disciplines (N = 97).

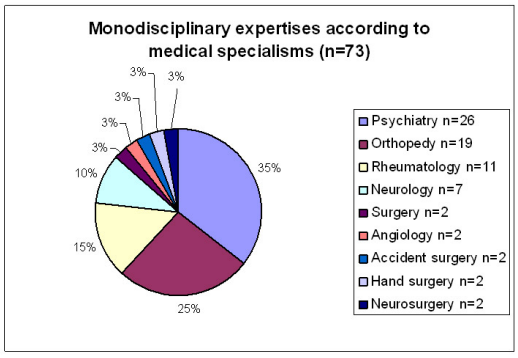

Figure 1c

Monodisciplinary expertises according to medical specialism (N = 73).

Plausibility analysis

As mentioned in the methods chapter, the aim of the plausibility analysis was to check the discrepancies regarding explicability between the reviewers.

Ninety three percent of the discrepancies in section I (formal aspects) could be cleared. Reasons for discrepancies in the assessment could, for example, be misunderstanding of a question or some points not being placed where reviewers expected them to be (for example, the history dispersed in the text). In the sections II and III (content questions) the discrepancies were reduced by 45% and 76% respectively. The explanations for these apparently high clearing rates in the plausibility checks will be dealt with in the previously mentioned separate report on the questionnaire.

Section IV (degree of difficulty) showed few discrepancies only (N = 10) which were not reduced by the plausibility procedure. The plausibility checks in section V (final overall grading) only concerned obvious mistakes caused by the reviewers or the asim team that were adjusted.

Analysis according to the final grading (section V)

In the overall quality rating 19.6% (95% confidence interval: 13.1%; 28.3%) of all the expertises were rated as insufficient. Figure 2 shows the cumulative distribution of the final grades with the two ratings on the same expertise connected. The grading of the two ratings on the same expertise deviated three times by more than one grade and once

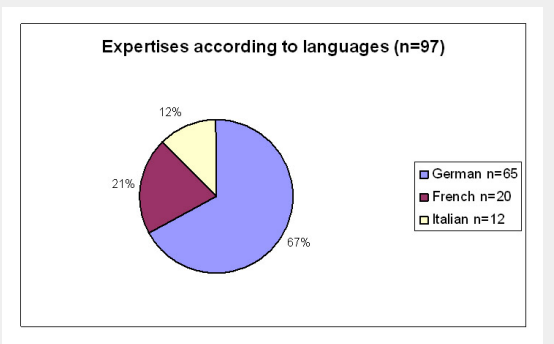

Figure 1d

Expertises according to languages (N = 97).

by three grades thus representing discrepancies, all 93 other rating pairs were identical or deviated by only one grade.

Influences on the final grading of expertises

The ***type of the expertises (mono-, bi- or polydisciplinary)*** had a significant influence on the proportion of expertises rated as insufficient (p-value: <0.01; fig. 3). The estimated proportion of insufficient expertises in monodisciplinary expertises was 0.12, in polydisiplinary 0.29, and in bidisciplinary 0.54.These results indicate that the monodisciplinary were rated better than the bi- or polydisciplinary expertises.

The ***language*** of an expertise had a significant influence on the proportion of expertises rated as insufficient (p-value: 0.019; fig. 3). The estimated proportion of insufficient expertises written in French was 0.02, in Italian 0.08, and in German 0.27. Thus expertises from the German speaking areas scored worse than those from the French and Italian speaking regions. Thirty five of the insufficient ratings originated from the German speaking part of Switzerland (N = 130 ratings), one rating from the French speaking part (N = 40 ratings) and two ratings from the Italian speaking part (N = 24 ratings).

The **duration of the impairment** had a significant influence whether an expertise was rated as sufficient or insufficient (OR: 1.13 [1.01;1.28], unit: year, p-value: 0.038). The longer the duration of the impairment the higher the proportion of insufficient expertises. The mean duration of the impairment was three years and 177 days, ranging from 58 days to 17 years. Whereas the 23 expertises (46 ratings) with less than one year of impairment were rated 44 times as sufficient and only twice as insufficient, the 74 expertises (148 ratings) with more than one year of impairment were rated 112 times as sufficient and 36 times as insufficient.

The judgement of the **duration of the processing** (period between the ordering of the expertise until its receipt) by the customer (only assessable expertises) had no significant influence on the expertise being rated as sufficient or insufficient (OR: 1.45 [0.53; 3.96], unit: year, p-value: 0.47).This means a longer duration of the processing does not necessarily coincide with a good (or poor) quality of the expertise.

The ***area of expertise*** (loss of income, liability, accident, disability insurance) had no significant influence on the proportion of expertises being rated as sufficient or insufficient (p-value: 0.66, fig. 3).

We found 12 insufficient ratings in a total of 50 ratings concerning disability, 17 in a total of 76 ratings concerning accidents, 3 in a total 18 ratings concerning liability and six insufficient ratings in a total of 50 ratings concerning loss of income.

The ***information*** upon which the expertise was based (records or patient contacts) had no significant influence on the proportion of expertises rated as sufficient or insufficient (p-value: 0.43; fig. 3).

For the monodisciplinary expertises the ***involved speciality*** did not significantly influence the proportion of the expertises receiving the rating sufficient or insufficient (p-value: 0.16, fig. 3).

The lack of a significant association between speciality and quality could be due to the substantial spread of the rating of the "neurological" cases and the spread of "other speciality" and "rheumatology". It is, however, of interest that in the neurological cases the proportion of insufficient expertises is close to 0.5, while for example, in the psychiatric cases the proportion is close to 0.

The assessment of the ***degree of difficulty*** was done according to a rating "A"-"E" ("A" meaning simple, "E" meaning very difficult, extremely complex). The individual reviewers have given the following ratings; rating "A" 30 times, rating "B" 101 times, rating "C" 39 times, rating "D" 19 times and rating "E" 5 times.

The assessment by the two reviewers was identical in 45 expertises, 1 grade different only and not exceeding the limit of sufficiency in 42 and discrepant in 10.

The influence of the degree of difficulty on the overall quality of the expertise could not be estimated because all expertises that were rated as "A" or "D" were consistently rated as sufficient, causing singularities and preventing proper modelling. Omitting ratings "A" and "D" from the analysis showed no significant association between difficulty and the proportion of insufficient expertises (p-value: 0.31, fig. 3). Nevertheless, a supportive analysis where the final rating (1–6) was treated as a continuous predictor (as it is frequently done with school grades) revealed a significant influence of the difficulty on the final ratings (p-value <0.001).

Association between the quality assessment by the clients in the comprehensive medical expertises study (MGS) and the quality rating in the current study

In the MGS study the ***comprehensibility*** of all expertises was rated as "substantiated" or "not well substantiated". This rating was significantly associated with the proportion of insufficient ratings on the overall quality by the reviewers (p-value: <0.01, fig. 4).

It is remarkable that ten of 16 (62.5%) expertises judged by both asim reviewers as insufficient were rated as sufficient by the clients. On the other hand, from only six of 69 (8.7%) expertises judged as "sufficient" by both reviewers were viewed as insufficient by the clients. Putting it in another way: when the asim reviewers judged the expertises as insufficient, only about one third of the cases were also rated insufficient by the clients.

Discussion

From a total of 3165 expertises in the MGS-study a stratified random sample of 97 expertises was taken, representing all areas of insurances. Although the random sample has been taken from a large number of expertises estimated to represent about a third of all Swiss medical expertises in the recruiting period, the investigation design of the quality analysis as such still has the characteristics of an extended pilot study as no validated evaluation instrument was available for use in this project.

The main part (N = 73) concerned monodisciplinary expertises. The involved specialities were mainly psychiatry, orthopaedics, rheumatology and neurology. In accordance with the linguistic regions in Switzerland, two thirds of all

expertises were in German language, the other expertises were in French or Italian. The 97 expertises were all analysed by two reviewers on the basis of a questionnaire consisting of five sections.

The present report describes the overall quality assessment (section V), the factors influencing the quality outcome and the evaluation procedure.

Quality of the expertises

On the basis of the results of this study, showing that 19.6% of the ratings are judged to be insufficient, the professional quality of medical expertises in Switzerland may appear to be unsatisfactory. However this observation needs confirmation by further research involving larger samples. Our data does not contradict the sceptical comments reported in the press [3]. It is remarkable that, in spite of the critical statements about the quality of the expertises and the time elapsed after the publications of the SUVA (2006), no essential improvement seems to have occurred.

In contrast to these earlier reports our assessment largely concentrates on professional, content related quality. However, in analysing section I of our questionnaire, we also looked at formal criteria as defined in the decision

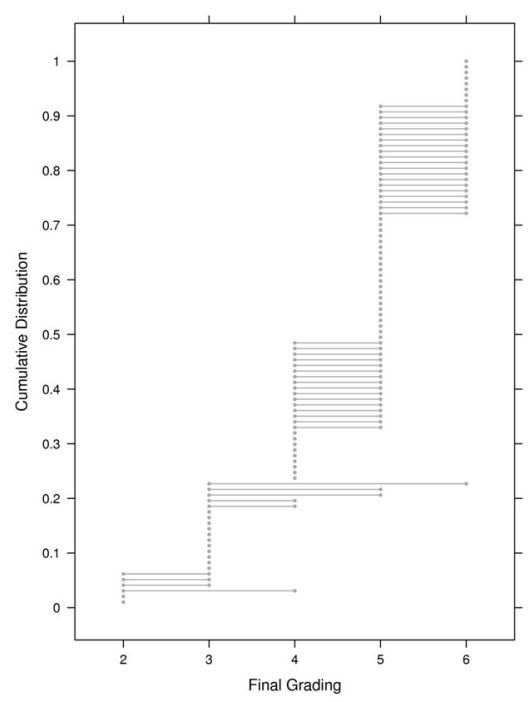

Figure 2

Final grading of all expertises (Section V).

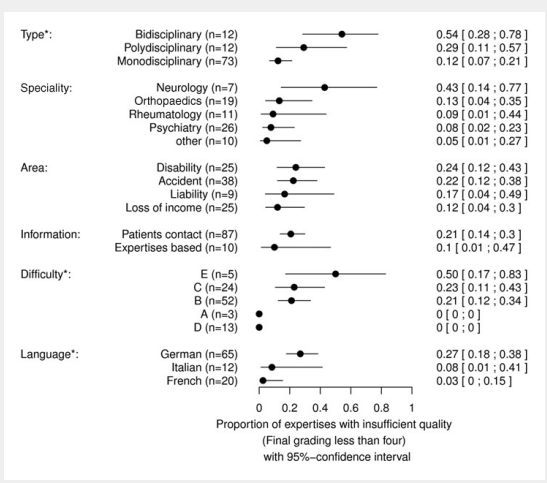

Figure 3

Influence of studied factors on the rating of overall quality. Factors that had a significant influence (type, difficulty, and language) are marked with a star. The number of expertises (n) in each category is indicated as well as the estimate and its 95% confidence interval.

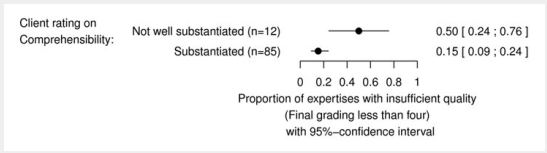

Figure 4

Association of the comprehensibility rating in the MGS-study ("substantiated", "not well substantiated") and the overall quality rating: The comprehensibility rating in the MGS-study was significantly associated with the proportion of expertises with insufficient quality (p-value <0.01). The number of expertises (n) in each category is indicated as well as the estimate and its 95% confidence interval.

Table 2: Allocation of expertises in the MGS-collective and the study sample.

	MGS-Collective (N = 3165)	**Study sample**(N = 97)
Kind of insurance		
Liability	38 (1.2%)	9 (9.3%)
Loss of income protection insurance	325 (10.2%)	25 (25.8%)
Disability insurance	2444 (77.2%)	25 (25.8%)
Accident Insurance	320 (10.1%)	38 (39.2%)
Others	38 (1.2%)	0 (0%)
Number of specialisms		
Monodisciplinary	2072 (65.5%)	73 (75.2%)
Bidisciplinary	437 (13.8%)	12 (12.3%)
Polydisciplinary	656 (20.7%)	12 (12.3%)
Languages		
German	2299 (72.6%)	65 (67.0%)
French	576 (18.2%)	20 (20.6%)
Italian	290 (9.2%)	12 (12.3%)

of the Swiss Federal Court (BG-judgment125 V 351) [1]. Nevertheless, the deficiencies observed were mainly professional content related issues. A more thorough and detailed analysis of the content sections II and III of the questionnaire and of expertises assessed as good or poor might lead to clarification in this respect. It will be the subject of a further report as previously mentioned.

The quality of the expertises was significantly influenced by the type of the expertises – mono-, bi- or polydisciplinary– and by the language. The monodisciplinary expertises appear to score better than bi- or polydisciplinary expertises. It may be that the integration step necessary to assemble the different assessments of bi- and polydisciplinary expertises into one expertise was an additional source of difficulties. Expertises in German were significantly worse than those in the Roman languages. A possible explanation, as seen by the board of reviewers, could be the higher specialisation and the ensuing training of the French and Italian experts. A more thorough analysis seems appropriate here, in order to rate, for example, the comparability of the expertise material in the different language groups.

The degree of difficulty had an influence on the overall assessment. The expertises that were classified as more difficult by the reviewers had better ratings. A possible explanation could be that potentially difficult expertises are assigned to specialised and qualified experts. The extent to which the mere rating as a difficult expertise influences the overall assessment needs to be investigated in more depth.

The duration from the moment of the impairment until commission of the expertise (duration of the impairment) had an influence on the rating. The longer this period lasted, the worse the assessments of the expertises became.

In comparing the estimates of the quality by the clients with the those of the present study, there appears to be a significant similarity regarding the motivation of the conclusions. Nevertheless the spread and the analysis of the sufficient and insufficient expertises varied strikingly. There were expertises consistently rated as sufficient in our study by both experts, that were judged as insufficient by the clients (6 from 69, 8.7%), however far more were judged reversely (10 from 16, 62.5%). Evidently, with all existing similarity, the quality criteria do not seem congruent a priori.

Questionnaire for the quality analysis

Though the detailed analysis of the questionnaire will be outlined elsewhere a few aspects deserve to be described in the present context. The questionnaire – a so far not validated tool – gave rise to considerable discrepancies between both reviewers, especially in section I (form) and III (argumentation). In section III school and the individual strictness could have mattered. The many contradictory answers to the formal questions in section I however must be due to different reasons, such as: unclear and inappropriate questions (e.g. for expertises based on records), misunderstanding of questions (clinical examinations with psychiatric expertises), unexpected structure of the expertise (that what is being searched for was not found in the usual place) and the difficulty to define the comprehensibility of the language. Understandably, many discrepancies could be cleared quite easily by an accurate analysis of the expertises by the asim team. In order to avoid such contradictory answers as much as possible, improvement remains mandatory. The discrepancies were definitely also due to the completely variable design of the expertises. Although mostly structured in a certain way, the structure was seldom standardised. It was therefore difficult for the reviewer to locate the single answers to the questions in the questionnaire.

An “unité de doctrine” in the composition of an expertise, as conveyed in expertise courses, would be helpful. Thereby all the relevant points in an expertise can be taken into account. Our questionnaire could therefore successfully be applied to expertises composed according to a clear scheme. It proved more difficult in the use for expertises with no clear or an individual design. It was totally unsuitable in certain parts (section I) for expertises based on records.

It became apparent that there were distinct differences between the Swiss and the German psychiatry reviewers in the application of the questionnaire. Partly, also in the somatic disciplines these differences between the reviewers became apparent.

Conclusion

The quality of expertises should be improved by professional postgraduate training. After all, 19.6% insufficient ratings in the selected sample of expertises ultimately means an unbearable burden for the patients and the insurers. The causes of the discrepancies between the professional content related quality and the assessment of the quality by the clients should be evaluated in more detail, especially with regard to loss of income protection insurance. A suitable quality reference for both clients and for experts is urgently needed. Based on the analysis of the discrepancies, the assessment instrument we have been using needs to be revised and adapted. The suitability of such an instrument should be tested in a further pilot study.

The significant variation in assessment of the quality between the expertises in German and those in French and Italian deserve to be further investigated. Possible contributing hypotheses that should be analysed are: perhaps better qualified colleagues perform the expertises in the Roman and Latin parts of Switzerland or the questions by the clients are better formulated, etc.

With a view to the permanent dissenting pattern of the final assessments (section V) and particularly of the sections covering the formal and content aspects (sections I–III), a standardisation of the assessment criteria is needed. This applies to the particular disciplines as well as to agreements that need to be arranged between the different schools of thought. The national and international professional dialogue regarding this topic should be ongoing.

Strength and weakness of the study

The strength of this study is that it uses a blinded analysis (medical assessor, patient, insurance company unknown to reviewer and vice versa) of expertises taken at random from a large and representative group of Swiss medical expertises covering all insurance areas and executed by an independent group of international experts. In addition, all the expertises were evaluated by two reviewers and for this purpose a special assessment instrument was developed.

Points of *weakness* are that the study still has pilot character, the assessment instrument concerned is a questionnaire that has not been previously validated, the sample size was small and there were many discrepancies between the reviewers recorded. Altogether, the study nevertheless conveys a reliable view of the quality of expertises in Switzerland, it however needs confirmation by future studies using larger samples and a validated evaluation tool.

The study was made possible thanks to the cooperation of our experts and reviewers:
Expert team:
André Aeschlimann, Jan Benthien, Heiner Bucher, Hans Ulrich Fisch, Dieter Frey, Harald Gündel, Paul Hasler, Peter Henningsen, Ralph Mager, Hermès Miozzari, Marco Mumenthaler, Pietro Regazzoni, Andreas Steck, Hans Rudolf Stöckli, Peter Villiger

Additional reviewers and consultants:
Sylvain-Frédéric Berner, Hans Peter Enderli, Andreas Hoffmann, Kurt Jäger, Liliana Mornaghini, René Raggenbass, Urs von Wartburg, Patrick Simon

We are grateful to our sponsors for their financial support:
Schweizerische Unfallversicherungsanstalt (SUVA), Bundesamt für Sozialversicherungen (BSV), IV-Stellenkonferenz (IV-SK), Axa Winterthur, SVV (Schweizerischer Versicherungsverband), SIM (Swiss Insurance Medicine), Consuldoc, Winterthurer Institut für Gesundheitsökonomie (WIG), asim (Academy of Swiss Insurance Medicine)

We are especially grateful to Pjotr Israels, the Netherlands, for the English translation of the manuscript.

Study funding/potential competing interests

The study was sponsored by Schweizerische Unfallversicherungsanstalt (SUVA), Bundesamt für Sozialversicherungen (BSV), IV-Stellenkonferenz (IV-SK), Axa Winterthur, SVV (Schweizerischer Versicherungsverband), SIM (Swiss Insurance Medicine), Consuldoc, Winterthurer Institut für Gesundheitsökonomie (WIG), asim (Academy of Swiss Insurance Medicine). No other potential conflict of interest relevant to this article was reported.

References

1 Bundesgericht, BGE 125 V 351. 1999.

2 Ludwig CA. Anforderungen an Gutachten – Anforderungen an Gutachter. SAeZ. 2006;87(23):1035–6.

3 Ludwig CA. Gutachtenqualität im Unfallversicherungsbereich. Medizinische Mitteilungen der SUVA, 2006. 77.

4 Schweizerische Gesellschaft für Rheumatologie – Arbeitsgruppe Versicherungsmedizin: Leitlinien für die Begutachtung rheumatologischer Erkrankungen und Unfallfolgen. SAeZ. 2007;88(17):736–42.

5 Mumenthaler M. Grundsätzliches zum ärztlichen Unfallgutachten. SAeZ. 2001;82(28):1521–4.

6 Schweizerische Gesellschaft für Versicherungspsychiatrie, Schweizerische Vereinigung ärztlicher Gutachter in Versicherungsfragen bei psychischen und psychosomatischen Störungen: Leitlinien der Schweizerischen Gesellschaft für Versicherungspsychiatrie für die Begutachtung psychischer Störungen. SAeZ. 2004;85(20):1048–51.

7 Fredenhagen H. Das ärztliche Gutachten. Leitfaden für die Begutachtung im Rahmen der sozialen und privaten Unfall-, Kranken- und Rentenversicherung. 2003, Bern: Verlag Hans Huber.

8 Riemer-Kafka G, et al. Versicherungsmedizinische Gutachten. 2007: Schweizerischer Ärzteverlag in Kooperation mit Stämpfli Verlag AG Bern.

9 Siegel AM, Fischer D. Die neurologische Begutachtung. 2004, Zürich: Orell Füssli Verlag.

10 AWMF. Arbeitsgemeinschaft der wissenschaftlichen medizinischen Fachgesellschaften, Uni Düsseldorf, Leitlinien für die Begutachtung von Schmerzen, AWMF-Leitlinien-Register, Nr. 030/102, Erstellungsdatum 10/2005, letzte Überarbeitung 03/2007 und Arbeitsgemeinschaft der wissenschaftlichen medizinischen Fachgesellschaften, Uni Düsseldorf, Ärztliche Begutachtung in der Psychosomatik und Psychotherapeutischen Medizin – Sozialrechtsfragen, AWMF Leitlinien-Register, Nr. 051/022, Erstellungsdatum 12.2.2001.

11 Deutsche Rentenversicherung. Qualitätsanalyse des ärztlichen Gutachtens, Prüfbogen der deutschen Rentenversicherung.

12 Ott W, Bögli M. Das medizinische Gutachten in HAVE. Personen-Schadenforum 2006, Tagungsbeiträge, Basel, 2006.

13 Auerbach H, Bollag Y, Eichler K, Gyr N, Imhof D, Stöhr S. Medizinische Gutachtensituation in der Schweiz, Studie zur Einschätzung der Marktsituation und zur Schaffung von Markttransparenz und Qualitätssicherung.

14 R Development Core Team. R; A Language and Environment for Statistical Computing. R. Foundation for Statistical Computing, Vienna, Austria. 2008.

15 Yan J, Fine J. Estimating equations for association structures. Statistics in Medicine. 2004.