ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-683

Application and Evaluation of a Medical Knowledge System in Sonography (SONOCONSULT)

Frank Puppe[1] and **Martin Atzmueller**[2] and **Georg Buscher**[3]
and **Matthias Huettig**[4] and **Hardi Luehrs**[5] and **Hans-Peter Buscher**[6]

Abstract. This paper presents the knowledge-system SONOCONSULT– an intelligent system in the medical domain. We evaluated the accuracy, acceptance and impact of SONOCONSULT, which has been used in clinical routine since 2002. The system was well accepted and had a significant clinical impact. In contrast to our original expectations, the diagnostic conclusions, although inferred with high accuracy, were less important especially for experienced physicians.

1 Introduction

Knowledge-based systems in medicine may serve many functions. Traditionally the main focus was on complex diagnostic and therapeutic recommendations [7, 4]. Recent reports [3] indicate that this may not be perceived as the primary need by most physicians. Instead, other functions such as support for high quality documentation, reminders, statistical analysis and training of beginners might be more important in clinical routine. We implemented a multifunctional knowledge-based system for sonography, which has been in routine use since 2002 documenting more than 12000 patients in two clinics, and evaluated its accuracy, acceptance and clinical impact. Table 1 shows a survey of the performed evaluations.

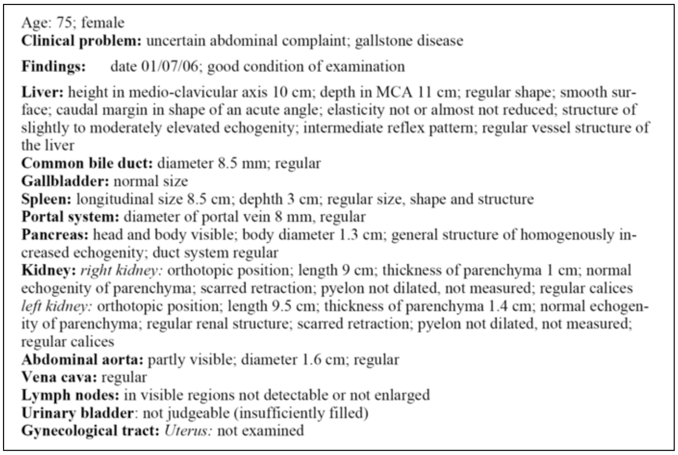
Evaluation	Procedure	Result
Diagnostic accuracy	99 prospective cases	High accuracy
Acceptance: Comparison of expectations and experiences	Interrogation of 14 resp. 19 physicians	High acceptance
Clinical impact: More complete reports?	103 reports	More complete
Internal consistency of reports	112 reports	Relevant discrepancy
Statistical interexaminer comparison	4100 cases	Relevant differences

Table 1. Survey of the performed evaluations.

SONOCONSULT (SC) [9] covers the entire field of abdominal ultrasound (liver, portal tract, gallbladder, spleen, kidneys, adrenal glands, pancreas, stomach, intestine, lymph nodes, abdominal aorta, cava inferior, prostate, and urinary bladder) and supports documentation, diagnosis, data mining and education. It was developed with the knowledge system d3web (www.d3web.de), which allows the input of expert knowledge via a graphical user interface [14].

The documentation component interacts with the user via dynamic questionnaires for all organs and generates two outputs: a structured report in a standard word processing system for the hospital information system and a data base of all cases for statistical analysis and data mining. The documentation component of SC has three modes (standard, short, and expert). The standard mode provides detailed and systematic questionnaires for all organs and is optimized by dialogue-guiding rules to ask only questions relevant for the case. The short mode asks questions on a more aggregate level, requiring more expertise from the user, and the expert mode allows directly entering just the diagnoses and measurements necessary for the report in a very compact manner. These documentation modes represent different compromises between the time necessary to document the case and the expertise of the user.

The terminology of SONOCONSULT is descriptive and follows that of standard textbooks and publications. Based on the completed questionnaires a textual report (see Figure 1) is generated using a rule based template. The report consists of three parts: 1) basic patient information, 2) findings and 3) judgement (which is added by the examiner as free-text). The SC-diagnoses are shown to the physician when entering the free text judgement, but they are not included in the report, because the physician remains responsible for the examination interpretation. The findings, judgements and diagnoses inferred by SC are also stored in a data base for statistical analysis.



Age: 75; female
Clinical problem: uncertain abdominal complaint; gallstone disease

Findings: date 01/07/06; good condition of examination

Liver: height in medio-clavicular axis 10 cm; depth in MCA 11 cm; regular shape; smooth surface; caudal margin in shape of an acute angle; elasticity not or almost not reduced; structure of slightly to moderately elevated echogenity; intermediate reflex pattern; regular vessel structure of the liver
Common bile duct: diameter 8.5 mm; regular
Gallbladder: normal size
Spleen: longitudinal size 8.5 cm; dephth 3 cm; regular size, shape and structure
Portal system: diameter of portal vein 8 mm, regular
Pancreas: head and body visible; body diameter 1.3 cm; general structure of homogenously increased echogenity; duct system regular
Kidney: *right kidney:* orthotopic position; length 9 cm; thickness of parenchyma 1 cm; normal echogenity of parenchyma; scarred retraction; pyelon not dilated, not measured; regular calices
left kidney: orthotopic position; length 9.5 cm; thickness of parenchyma 1.4 cm; normal echogenity of parenchyma; regular renal structure; scarred retraction; pyelon not dilated, not measured; regular calices
Abdominal aorta: partly visible; diameter 1.6 cm; regular
Vena cava: regular
Lymph nodes: in visible regions not detectable or not enlarged
Urinary bladder: not judgeable (insufficiently filled)
Gynecological tract: *Uterus:* not examined

Figure 1. Part of a generated exemplary SONOCONSULT-report.

The diagnostic component adds inferred diagnoses based on the input data from the questionnaires to the output. The knowledge base makes use of medical heuristics as a knowledge source [12] and was built according to the principles applied for the construction of HepatoConsult [5]. SC uses five main concepts: symptoms (input data), symptom classes (Questionnaires grouping the input questions), symptom abstractions, diagnoses (output), and rules. Symptoms consist of a pair (attribute, value), where the attribute is the

[1] University of Wuerzburg, puppe@informatik.uni-wuerzburg.de
[2] University of Wuerzburg, atzmueller@informatik.uni-wuerzburg.de
[3] DFKI, Georg.Buscher@dfki.uni-kl.de
[4] DRK-Kliniken Berlin-Koepenick, matthias.huettig@arcormail.de
[5] University-Hospital Wuerzburg, h.luehrs@medizin.uni-wuerzburg.de
[6] DRK-Kliniken Berlin-Koepenick, buscher.dhp@t-online.de

symptom name (e.g. liver size) and the value is the symptom value (e.g. increased). In interactive settings, the attributes are questions and the values are the answers by the user. There are two main types of attributes: choice and numerical. Choice attributes have a predefined range (e.g. for liver size: decreased, normal, increased) and are differentiated according to their cardinality as one-choice (1, i.e. exactly one value is allowed, like for liver size) or multiple choice (0 .. n). Symptoms are grouped into symptom classes if they are requested together most of the time. It is possible to define rules in a symptom class that specify which questions have to be asked in which order depending on the values of previously answered questions. Symptom abstractions are very similar to symptoms except that their values are inferred by rules. They allow a stepwise abstraction of the input data. Diagnoses are also inferred by rules from symptoms, symptom abstractions or other diagnoses (”criteria”). They usually aggregate uncertain evidence. While d3web allows different reasoning mechanisms for inferring diagnoses, in SC a score-based scheme is used, i.e. the rules are assumed to be independent and add or subtract points to the score of a diagnosis, which is rated by thresholds in one of the linguistic categories ”probable”, ”possible” and ”unclear or excluded”. Rules consist of a condition, an action and exceptions. The condition may be a nested logical combination of criteria, e.g., ”and”, ”or” and ”not”. Rule actions include, e.g., rating diagnoses, computing values for symptom abstractions, indicating symptom classes and (further) follow-up questions. Exceptions allow to differentiate between two types of negation, i.e., whether a fact is yet unknown or definitely wrong. For more details, see [14].

The diagnostic procedure of SC follows the hypothesis-and-test- and the establish-refine-strategy. The selection of a specific questionnaire (symptom class) depends on the overall clinical question and on the inferred diagnoses. Data gathering stops when (a) the user jumps to the conclusions or (b) all suspected diagnoses (category ”possible”) are either ”probable” or ”unclear or excluded” by means of the program's expertise or (c) there are no useful questionnaires left for clarification. Besides the 430 questions, SC contains about 140 symptom abstractions, 230 diagnoses and several thousand rules of varying complexity. In an average case, from the 60 entered symptoms 20 symptom abstractions and 3-6 diagnostic conclusions are inferred by the program. The range 3-6 means that some inferred diagnoses of SC are less important than others (like e.g. ”adiposity”) and could be ignored depending on the point of view.

The data mining component offers a standard tool for getting an overview of the data and an innovative subgroup analysis tool for knowledge discovery and quality control. The data mining technique of subgroup mining [10, 8] is quite suitable for common medical questions, e.g. whether a certain pathological state is significantly more frequent if combinations of other pathological states exist or if there are diagnoses and/or findings, which one physician documents significantly more or less often than the average. We used the VIKAMINE (Visual, Interactive and Knowledge-Intensive Analysis and Mining Environment) system [1] for interactive and automatic subgroup mining. This tool is adapted to particularities of the medical domain like many missing values in the records due to intelligent data gathering strategies minimizing the number of asked questions. Furthermore, often background knowledge can be utilized, since existing knowledge should not be rediscovered, but the available knowledge should be used to find new, often subtle correlations, to increase the interestingness of the discovered results. Additionally, often (known) confounding factors (like age, gender, body weight etc.) need to be controlled. VIKAMINE offers an efficient exhaustive and various heuristic search options with constraints for automatic subgroup discovery and interactive visualizations for active user involvement. When the user discovers something unexpected/interesting in the data using standard tools, then these findings can be inspected and analyzed in detail using VIKAMINE. For more details, see [1].

The educational effect of SC is based on the structured documentation procedure showing what aspects are important in what context, and the explanation component showing the diagnostic meaning of observations and vice versa the criteria for inferring a diagnosis. Additionally, most findings and diagnoses are linked to a text-book-like information system for rapid information lookup.

The rest of the paper is organized as follows: In Section 2 we discuss the clinical experience with the system, evaluations of its accuracy, acceptance, and clinical impact, and discuss these results. Section 3 describes an application of the data mining component for the detailed analysis of interexaminer variations. In Section 4 we present the lessons learned, and conclude the paper in Section 5 with a discussion of the presented work and promising options for future work.

2 Clinical Experience and Evaluations

SC has been in routine use since 2002 as the only documentation system for ultrasound examinations in the DRK-hospital of Berlin-Köpenick; since 2005, it is in routine use at the university hospital of Würzburg. Since SC runs on a web server, intranet integration with the hospital information system (HIS) was a prerequisite for both clinics. In Berlin a weak integration was implemented: the physician uses two separate programs: SC for entering the sonographic data and the HIS for storing the document generated by SC. The transfer of the data is done in a ”copy and paste” style, that is implemented with a ”one-click”-macro for convenience. In Würzburg, a strong integration was implemented: The physician starts SC from the HIS, where SC is initialized with some basic patient data. After finishing the case, SC transfers the report into the physician's letter section of the HIS and the ICD-coded diagnoses in the diagnostic section of the HIS. In Berlin the standard documentation mode of SC is used, whereas in Würzburg the expert mode is applied. We evaluated the diagnostic accuracy of SC, the clinical acceptance and the clinical impact. Due to the longer period of routine use, most of the following evaluations were done in Berlin-Köpenick.

2.1 Accuracy

We define diagnostic accuracy as consistency between the input and output data, i.e. whether the inferred diagnoses are consistent with the entered symptoms. The standard documentation mode is then most appropriate, since the short and the expert mode require the physician to enter own interpretations not really challenging the diagnostic power of SC. We used 99 consecutive cases from Berlin in a prospective study. As gold standard, one sonographic expert from a different clinic than Berlin performed the evaluation, since the goal was to show that the SC knowledge base did not contain serious errors. The evaluator got the findings including the clinical problem and the list of diagnoses SC inferred from the findings. Rating each diagnosis, the evaluator entered the overall impression of the case using four categories (SC-diagnoses fully consistent with findings, basically consistent, partly deviating and seriously deviating).

After the evaluator had completed the forms, the developer of the knowledge base of SC classified each error with the following categories: a) Judgement difference, e.g. due to different thresholds used by the evaluator and SC for organ size, b) input error, i.e. the documented findings are inconsistent leading to erroneous conclusions, and c) knowledge base error due to either rule errors or errors in the template applied for generating the text. According to the overall impression, 92% of the cases were rated correct or basically correct (i.e. the diagnoses were consistent with the documented findings), in 7% of the cases, the diagnoses of SC were partly deviating and in 1% were seriously deviating from the documented findings (in this case the documented findings were inconsistent). A closer look on the level of the individual diagnoses showed that 83.9% of all 483 diagnoses in the 99 cases were rated consistent. If only important diagnoses were considered, even 94.9% of diagnoses were rated as consistent. The classification of the errors showed that most errors were due to judgement differences between the developer of SC and the evaluator. This is not surprising, because judgement in sonographic examinations is in part subjective. In particular, different thresholds for normal organ sizes were responsible for the majority of expert disagreements. However, if only important diagnoses or the overall rating of the cases are considered, the judgement differences are less prominent and the main reasons are input errors. Knowledge base errors were responsible only for 2% of cases with partly deviating conclusions and 0.7% of wrong diagnoses.

2.2 Acceptance

The acceptance of SC in Berlin was measured with a before-after comparison, i.e. the users were asked to fill out a questionnaire before SC was installed in routine use and to fill out a second questionnaire two years after its installation. According to the users opinion, the most important preconditions for the programs introduction into clinical routine were (a) an acceptable account of symptom representation, (b) a time-efficient input procedure, and (c) the ability to convert the case data into structured text documents for the medical record of the procedure. These preconditions were met before the program was put into routine use. While a self written report took on average about 5 minutes for senior examiners to complete an examination using a text system including some building blocks for common phrases, the input time with SC was about 4-13 minutes (mean 7.55 minutes) when starting to work with the program and about 5 minutes after being familiar with it for about 2-3 weeks of continuous use in the standard documentation mode. The expectations of the prospective users of SC were queried prior to its first presentation. We provided a questionnaire that was answered by 19 sonographic examiners. After gaining experience with the use of SC, the physicians were asked again about their opinions using a questionnaire that was answered by 14 examiners. Both questionnaires asked items similar to a five point Likert scale.

The answers to these questions show that prior expectations (PE) and the actual experiences (AE) agree in many aspects: the standardization of nomenclature and subsequent comparability of sonographic reports is acknowledged by the examiners (PE: 4.3; AE: 4.5), simple usability is very important (PE: 4.9; AE: 3.8) and the reminder function of the program is perceived as helpful (PE: 3.7; AE: 3.8). This is also true for the presentation of the system diagnoses, which is perceived as not so important (AE: 3.0; PE: 2.9); the influence of the system diagnoses on the diagnoses of the physicians was rated even lower (PE: 2.2). A difference between expectations and experiences exists with respect to the explanation function, which was declared as rather desirable, but rarely used (PE: 3.8; AE: 2.5). The expected training effect (PE: 3.9) was compared with the experiences of 5 beginners and clearly confirmed the expectations. They all emphasized that the program's most positive effect was to conduct an examination in a complete and structured way as well as in a standardized and reasonable sequence. The diagnostic properties of the program had only been of medium/transitory interest during the learning phase.

2.3 Clinical Impact

We also tried to measure whether the use of SC improved the quality of the sonographic records: Potential improvements are a more complete documentation of symptoms and a higher quality of the reported diagnoses. Concerning the first issue, after the introduction of SC in Berlin the program established a documentation standard, which is highly welcomed by the physicians (see evaluation of acceptance). The question was how complete the sonographic reports would have been without applying SC. Therefore, we randomly selected 103 hand written reports, which were documented before the introduction of SC in Berlin and noted whether all questions asked by SC could be answered with the available data. If not, two senior examiners from the clinic in Berlin judged the information gaps in the free text reports as relevant or dispensable. The evaluation showed that 287 information gaps were found (i.e. questions generated by SC which could not be answered considering the sonographic reports); the domain experts judged nearly half of them (132) as relevant.

To confirm the assumption of a documentation standard after the introduction of SC in the first study and for evaluating the second issue concerning the quality of the diagnoses, we performed another study: We used 112 prospective consecutive records and judged the completeness of the documented findings and the consistency of the diagnostic conclusions with the documented findings. The agreement of three domain experts (2 from the clinic in Berlin and one from Würzburg) was used as ”gold standard”, i.e. the diagnostic conclusions were judged by the domain experts as ”correct” or ”problematic”, when at least two agreed on the same assessment. The evaluation confirmed that there were little information gaps in the reports, i.e. the examiners had answered nearly all the questions SC asked them. From 412 ”true” diagnoses in these records (i.e. in this sample an average of 3.7 per case), the examiners missed 107 (26%) diagnoses in their free text judgement and stated an additional 32 diagnoses, which were not supported by the documented findings. The evaluators also informally rated the diagnostic conclusions of SC. In agreement with the accuracy evaluation mentioned above, the SC-diagnoses were judged in general as adequate by the evaluators. The difference between the consistency of diagnoses of SC and the examiners was unexpected, because the examiners were shown the diagnostic conclusion of SC before entering their free text judgement.

We differentiated the 412 diagnoses further into simple and complex conclusions (the latter are based on the combination of more than one symptom). There were 145 complex diagnoses, from which the examiners missed 57 (39%). The examiners stated 15 additional complex diagnoses unsupported by the documented findings. That means, the inconsistency between findings and diagnoses was higher for complex diagnoses (39% compared to 26%). These surprising figures are difficult to interpret with respect to the clinical correctness of the diagnoses, since the evaluation was based on text documents, not on sonographic pictures, because these were not included in the records. Therefore, in general it is not possible to differentiate between incorrect symptom descriptions and incorrect conclusions, although the relative high degree of problematic simple diagnoses (50 from 267, i.e. 19%) indicates some documentation errors. Nevertheless, the remaining inconsistency between documented findings and diagnostic conclusions is rather high. As mentioned above, this is quite astonishing, since the SC-diagnoses were visible to the examiners before writing their final comment: It is questionable whether they considered the diagnostic SC-conclusions for verification of their data input. This fact is consistent with the low influence of the system's diagnoses on the own diagnoses of the examiners (see 2.2).

To investigate this phenomenon further, the silver bullet would be a study where the quality of sonographic reports is judged by comparing the judgements of the examiners with those of independent experienced sonographic experts, who examine the same patients a second time. Even using pictures from the first examination to be judged by experts instead of a second examination might cause a bias. Since this procedure must be repeated several times for different uses of SC and different examiners, we considered it as too expensive for our routine evaluations. Instead, we focused on statistical quality control as presented in the next section.

3 Statistical Analysis

The physicians considered statistical analysis as one of the desirable features. About 300 detailed patient records are documented per month in each clinic in Würzburg and Berlin. Statistical analysis ranges from getting an overview on the relative frequency of sonographic diagnoses via detecting patterns specific to different examiners and their experience and correlations of sonographic diagnoses with final clinical diagnoses to knowledge discovery of correlations among pathological states of different organs and risk factor analysis. We used the subgroup discovery tool VIKAMINE for interactive analysis and focus in the following on the analysis of interexaminer variations in Berlin (see [2] for methodological issues).

In Berlin the examiners rotate according to a predefined schedule, e.g. every six month. We used for this study sonographic data over a period of 3 years with 7 different young examiners (E1 ...E7), and considered the first 600 consecutive cases from each examiner. We checked that the case mix of the different examiners with respect to age group and gender was roughly the same. In a first step we analyzed the distribution of all diagnoses, and found considerable variations between the 7 examiners for several diagnoses: We discuss two examples: liver cirrhosis and chronic renal failure (CRF).

Examiner E1 rated the diagnoses CRF as probable or possible more than ten times as often as E5 and twice as often as the average. The detailed analysis of E1 revealed that only one of two possible parameters is responsible for this special rating: a narrowed left or right renal parenchyma, while e.g. the renal size is not significantly different from the average. Examiner E3 rated the diagnosis liver cirrhosis as probable or possible more than four times as often as E7 and more than twice as often as the average. The detailed analysis of E3 showed that the combination of the findings ”rarefaction of portal branches” and ”liver plasticity moderately reduced” is responsible for this increase, since E3 has a share of 90% of this combination (compared to a share of 14% for all cases), from which liver cirrhosis = possible is inferred in nearly all cases.

This analysis shows that the differences between examiners in rating diagnoses can be traced back to one or two specific findings, which offers the opportunity to focused training actions in clinical routine. The use of the standardized documentation system SONOCONSULT offers the opportunity to link additional informal knowledge (e.g., pictures) for differentiation of critical findings, which are just a mouse click away during documentation. It also seems worthwhile to offer the examiners the possibility to compare their examination profiles computed by the subgroup mining tool to be informed about deviations, which might trigger a look on the respective informal knowledge. Further, the diagnostic results of SC can be used as motivation for reevaluation of the data input as well as of the diagnostic conclusions. This necessitates a simple presentation of the – with respect to the conclusions – incongruent inputs. Finally, SC in combination with data mining may be used to generate individualized quizzes with multiple choice questions which can be solved online by young sonographic examiners in training on a regular base. The effect of such clinical actions can then be evaluated in prospective studies using the same subgroup mining tools as described above.

In summary, the results indicate a high variability of documentation and interpretation habits of the different examiners. This kind of statistical quality control indicates that different examiners vary in their performance. This observation is in line with the noted inconsistencies with respect to documented findings and inferred diagnoses in section 2.3. A possible interpretation is that the quality of sonographic reports depends much on the individual skills of the examiners, i.e. examination experience and accuracy of documentation. Especially the interpretation that the documented findings are less reliable than the diagnostic conclusions of the examiners is of interest and should be a focus of further development of control instruments. However, it must be taken into account that the set of patients investigated by the different examiners might have different characteristics. Along the line of investigating such hypotheses with statistical means, we plan further studies, where we compare the sonographic diagnoses with the final diagnoses as stated in the coded diagnoses of the hospital information system or the (informal) diagnoses mentioned in the physician's letter. Additional hints can be derived from comparison of sonographic diagnoses with diagnoses from lab data, computer tomography or magnet resonance imaging. However these studies – except using the coded diagnoses of the HIS – require a formalization and standardization of the diagnoses in these reports. This is a difficult and time-consuming task requiring computer based assistance, which we just started to do. Even coded ICD data requires a mapping, because first studies showed that the ICD codes from sonographic diagnoses were not identical with ICD codes for equivalent final diagnoses. Therefore results are currently not available.

4 Lessons Learned

Computer based documentation and diagnosis systems can increase the quality of documentation and diagnosis. In our five years of experience with SONOCONSULT we have shown some significant effects and have observed indications for other important effects. In particular, the use of intelligent dynamic questionnaires increases the completeness of records without prolonging the time necessary for data input. Since all records use the same terminology, they become comparable, which was welcomed in itself by the physicians. In addition, it enabled different forms of quality control. Our results show that there is some need for quality control, which has been undetected so far. In order to avoid expensive studies where sonographic examinations resp. interpretations had to be repeated by experts, we used computer-based methods for quality control:

- The documented findings and the conclusions of the examiners in the records are not always consistent. Rule-based computer-diagnosis achieves a much higher consistency.
- Different examiners have a relatively high variation in stating sonographic diagnoses like liver cirrhosis or chronic degenerative renal failure, due to variations in reporting corresponding findings.
- Automatic comparison of sonographic diagnoses with diagnoses from other data sources (lab data, CT, MR) and the final diagnoses may give further valuable hints for quality control. However, we just started to follow this path which seems quite promising.

A rather unexpected finding was that experienced examiners largely ignored computer generated diagnostic suggestions, although they were of high quality. This observation is supported asking the physicians about their attitudes and their actual behaviour when stating their diagnostic conclusions. Less experienced examiners welcomed the systematic approach and the diagnostic conclusions. It seems worthwhile to adopt the well-known critiquing approach [13] to draw the attention of examiners to inconsistencies in the report and simultaneously to allow them the fast correction of the inconclusive entries. This is in line with the observation in [6] that physicians often do not know when their diagnoses are incorrect.

At the Würzburg university medical hospital, SONOCONSULT was adopted two years ago, taking this observation into account. The system was used in a different (expert) mode: The physician first enters the diagnoses and the computer then asks about the most important findings necessary for inferring them. In this way, the time for entering data is reduced. The entered diagnoses are automatically ICD coded, and these codes are automatically transferred into the hospital information system as sonographic diagnoses. The system is still able to do some consistency checking, although much less than in the standard mode as used in Berlin. This mode is not suitable for beginners, but optimized for experienced examiners. Since the evaluation of the accuracy of SC showed that parts of the knowledge base should reflect different opinions (e.g. organ size thresholds), it was necessary to make these parts adaptable. This seems to be a general prerequisite for transfer from one clinic to another.

These lessons learned can be generalized to the insight that the GUI of an interactive knowledge system should be designed to integrate and complement the competence of (potential heterogenous) users instead of duplicating it, e.g. diagnostic knowledge can be used for beginners to infer diagnoses, but for experienced physicians, it is more acceptable to be used for semi-automatic report generation or maybe a critiquing approach.

We currently plan two obvious steps for quality improvement: First, to reduce variations among examiners, the questions of SC concerning the findings with high variations will be annotated with descriptions and in particular reference pictures. In addition, the examiners are informed, if they deviate considerably from the average frequency for certain diagnoses and/or the corresponding findings. Second, the observation that SC can infer more consistent diagnoses as the examiners from the documented findings can be used for a critiquing approach in Berlin. Thus, the diagnoses of the examiner are compared to the diagnoses inferred by SC, and if there are serious discrepancies, the examiner is informed about the inconsistency and offered means for correction (either to change the diagnoses or to change the underlying findings). The difficulty with this approach is to extract diagnoses from the free text judgement of the examiner. A suitable technique is expectation driven information extraction [11], which is quite promising, but currently not fully operational, because the quality of information extraction must be very high. The effects of quality improvement activities will be measured with statistical techniques as outlined above.

5 Conclusions

The applications and the evaluations of SONOCONSULT showed (1) its benefits as an intelligent documentation system producing more complete records in a standardized nomenclature in about the same amount of time as hand-written reports, (2) its training value for beginners, (3) its high diagnostic accuracy, and (4) its potential for statistical quality control. Although the system was well accepted in general, its diagnostic conclusions were largely ignored. We reacted in two ways: when migrating SONOCONSULT from Berlin to Würzburg, we defined a new mode of data entry with intelligent questionnaires, where the diagnoses are entered first and supporting findings are asked subsequently. This shortened the time for data entry but depends on the clinical knowledge of the examiner. Especially for beginners in sonography – a common problem due to planned rotations - the standard mode will be supported with a critique component, where the free text judgement of the examiner and the diagnostic conclusions of SONOCONSULT are compared and the examiner is informed about major discrepancies.

We also plan to compare the sonographic diagnoses systematically with diagnoses from other investigations (lab data, CT, MNR) and in particular the final clinical diagnoses. However, these projects require limited understanding of free text reports, which we address with the technique of expectation driven information extraction. The prospect is that we can evaluate the clinical significance of sonographic examinations in a systematic manner without relying on the expensive technique to recheck individual sonographic examinations manually by different experts, despite this might be the ultimate gold standard.

ACKNOWLEDGEMENTS

This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-1 and Pu 129/8-2.

REFERENCES

[1] Martin Atzmueller, *Knowledge-Intensive Subgroup Mining*, volume 307 of *Diski*, IOS Press, 2007.

[2] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher, 'Profiling Examiners using Intelligent Subgroup Mining', in *Proc. 10th Intl. Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005)*, pp. 46–51, Aberdeen, Scotland, (2005).

[3] *Clinical Decision Support Systems*, ed., E. Berner, Springer, 2007.

[4] Maisiak R. Cobbs G. Taunton O. Berner, E., 'Effects of a Decision Support System on Physicians' Diagnostic Performance', *J Am Med Inform Assoc*, **6**, 420–427, (1999).

[5] Hans-Peter Buscher, Ch. Engler, A. Führer, S. Kirschke, and F. Puppe, 'HepatoConsult: A Knowledge-Based Second Opinion and Documentation System', *Journal Artif. Intell. in Med.*, **24(3)**, 205–216, (2002).

[6] Timothy M. Franz Gwendolyn C. Murphy Fredric M. Wolf Paul S. Heckerling Paul L.Fine Thomas M. Miller Arthur S. Elstein Charles P. Friedman, Guido G. Gatti, 'Do Physicians know when their Diagnoses are correct? Implications for Decision support and Error Reduction', *J Gen Intern Med.*, **20**, 334339, (2005).

[7] Poynard T. Darmoni, S., 'Computer-Aided Decision Support in Hepatology', *Scand J Gastroenterol*, **27**, 889–896, (1992).

[8] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, USA, 2006.

[9] Matthias Huettig, Georg Buscher, Thomas Menzel, Wolfgang Scheppach, Frank Puppe, and Hans-Peter Buscher, 'A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography', *Medizinische Klinik*, **99**(3), 117–122, (2004).

[10] Willi Klösgen, *Handbook of Data Mining and Knowledge Discovery*, chapter 16.3: Subgroup Discovery, Oxford Univ. Press, NY, 2002.

[11] Schuhmann M. Puppe F. Buscher H.-P. Klügl, P., 'Expectation-Driven Information Extraction in Incomplete Sentences (in German)', in *Proc. of LWA 2007 in Halle, Germany*, pp. 237–243, (2007).

[12] Clement J. McDonald, 'Medical Heuristics: The Silent Adjudicators of Clinical Practice', *Ann. Intern. Med.*, **124**, 56–62, (1996).

[13] Perry L. Miller, *Expert Critiquing Systems - Practice-Based Consultation by Computer*, Springer Verlag, 1986.

[14] Frank Puppe, 'Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3', *Intl. Journal of Human-Computer Studies*, **49**, 627–649, (1998).