Clinical Experiences with a Knowledge-Based System in Sonography (SonoConsult)

Frank Puppe[1], Georg Buscher[1], Martin Atzmüller[1], Matthias Hüttig[2], Hans-Peter Buscher[2],
[1] Universität Würzburg, Lehrstuhl Informatik VI, Am Hubland, 97074 Würzburg
[2] DRK-Kliniken Berlin Köpenick, Medizinische Klinik 2, Salvador-Allende-Str. 2-8, 12559 Berlin

Abstract: We evaluated the clinical effects of the knowledge-based documentation and diagnosis system SonoConsult for sonography, which has been used in clinical routine for more than 2 years. The evaluation focuses on the following aspects from the clinical point of view: quality of documentation, quality of diagnostic conclusions, training effects, and research effects. In contrast to wide-spread expectations in the knowledge-based community, the diagnostic conclusions were less important than the other aspects, being much more welcomed by clinicians.

1 Introduction

Knowlegde based systems in medicine may serve many functions. Traditionally the main focus was on diagnostic and therapeutic recommendations [Darmoni et al. 92, Berner et al. 99]. However, this may not be perceived as the primary need by most physicians. Instead, other functions such as support for high quality documentation might be more important in clinical routine. We implemented a multifunctional knowledge-based system for sonography and evaluated both its acceptance and its clinical impact.

2 Structure and Function of SonoConsult

SonoConsult (SC) [Hüttig et al. 04] was developed with the diagnostic shell kit d3web [www.d3web.de] which allows the input of expert knowledge via a graphical user interface [Puppe 98]. It covers the complete field of abdominal ultrasound. The implementation makes use of medical heuristics as a knowledge source [McDonald 96] and was performed according to the principles of construction of HepatoConsult [Buscher et al. 02]. The web interface is adaptable for communication with clinical information systems. The terminology of symptoms is descriptive and follows that of textbooks and publications. The interpretation of symptoms with respect to intermediate conclusions and final diagnoses is made by weighting points (scores). Using thresholds, they are summarised and rated into the categories “probable”, “possible” and “unclear or excluded”.

The diagnostic procedure follows the hypothesis-and-test- and the establish-refine-strategy. The selection of a specific questionnaire depends on the overall clinical question and on the inferred diagnoses. Data gathering stops when all suspected diagnoses are either established or excluded by means of the program’s expertise. Suspected diagnoses which cannot be established or excluded are designated as “possible” diagnoses. Subsequently, the case record is stored in a data base. On the basis of the case record a structured text document (medical report) is generated using a predefined template. It consists of four parts: basic patient data, differentiated report of symptoms, system diagnoses (automatically inferred) and examiner comment (free text). This document is added to the patient record. The screenshot in Fig. 1 exemplifies the data input in SC und Fig. 2 shows the corresponding generated report.

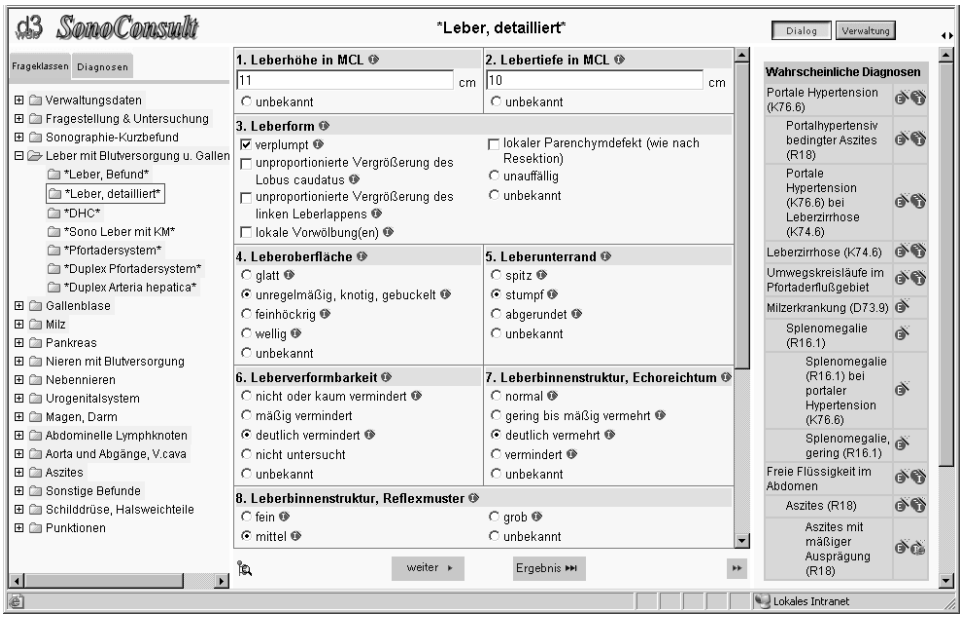

Figure 1: Screenshot of a section of an SC-questionnaire with part of the hierarchy of questionnaires (partially opened) exemplarily showing the degree of specification (left panel) and the currently generated probable system diagnoses (right panel).

The program includes an explanation tool enabling the user to retrace a diagnostic pathway of inferences from symptoms to diagnoses. Additionally, all symptoms and diagnoses are linked to a text-book-like information system for rapid information lookup.

SC was developed continuously on the basis of user feedback. It contains about 430 questions for symptoms, 140 symptom interpretations and 230 diagnoses. The analysis of the data base of 770 consecutive cases exhibited a mean of 61 questions per case with an average of 20 symptom interpretations and 6 diagnoses inferred by the program.

3 Clinical experience and evaluations

3.1 Acceptance

For more than two years SC is in routine use as standard documentation system for ultrasound examinations in the DRK-hospital of Berlin-Köpenick. According to the users' opinion, the most important preconditions for the program's 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 about 3-5 minutes for senior examiners, the input time for a complete case was about 5-8 minutes when starting to work with the program and 2-4 minutes after being familiar with it after about 2-3 weeks of continuous use.

Sonographie

Name, Vorname: **Mustermann, Manuel, 01.10.40**
Fragestellung: Oberbauch-Screening; Leberzirrhose

Befund vom 17.11.04; gute Untersuchungsbedingungen

Leber: Höhe in MCL 11 cm; Tiefe in MCL 10 cm; verplumpt; Oberfläche unregelmäßig, knotig, gebuckelt; Unterrand stumpf; Verformbarkeit deutlich vermindert; Binnenstruktur deutlich echovermehrt; mittleres Reflexmuster; Kalibersprung der Pfortaderäste intrahepatisch; Rarefizierung der Pfortaderäste intrahepatisch
D. hepatocholedochus: Durchmesser 5 mm; unauffällig
Gallenblase: unauffällig
Milz: längs 14 cm, tief 6 cm; Parenchym unauffällig
Pfortadersystem: Pfortaderdurchmesser 14 mm; keine wesentliche Zunahme des Durchmessers bei Inspiration; Milzvenendurchmesser 12 mm; Hinweis auf wiedereröffnete Nabelvene
Duplexsonographie: Pfortader Fluß orthograd mit gleichmäßigem Flußprofil, Flußgeschwindigkeit 12 cm/s; Milzvene Flußgeschwindigkeit 12 cm/s; wiedereröffnete Nabelvene
Flüssigkeit im Abdomen: freie Flüssigkeit im Sinne von Aszites, mäßig ausgeprägt
Abdominelle Gefäße: *Arteria hepatica (duplexsonographisch):* nicht durchgeführt
Vena cava: unauffällig
Lymphknoten: in beurteilbaren Regionen nicht erkennbar bzw. nicht vergrößert
Pleuraerguss: beidseits nicht nachweisbar
Perikarderguss: nicht nachweisbar

Beurteilung:
Schlussfolgerungen von SonoConsult:
Portale Hypertension (K76.6) bei Leberzirrhose (K74.6); portalhypertensiv bedingter Aszites (R18); Splenomegalie (R16.1) bei portaler Hypertension (K76.6)
Die diagnostischen Schlussfolgerungen müssen durch den Untersucher/Befunder geprüft werden.

Bemerkung des Befunders:
Leberzirrhose mit portalhypertensiv bedingtem Aszites und mäßiger Splenomegalie. Kontrolle nach Ausschwemmung des Aszites empfohlen.

Figure 2: Generated exemplary SC-report corresponding to the ultrasound examination in Fig. 1.

The expectations of the prospective users of SC were asked before its first presentation. We provided a questionnaire that was answered by 19 sonographic examiners:

A1 Influence on examination procedure: 3.1 (1 = not probable; 5 = highly probable)
A2 More time for documentation: 3.6 (1 = no willingness; 5 = high willingness)
A3 Standardized data input: 4.3 (1 = unimportant; 5 = very desirable)
A4 Indication of incomplete examination: 3.7 (1 = unimportant; 5 = very desirable)
A5 Presentation of system diagnoses: 3.0 (1 = unimportant; 5 = very desirable)
A6 Simple usability: 4.9 (1 = unimportant; 5 = very desirable)
A7 Rapid access to previous results: 4.6 (1 = unimportant; 5 = very desirable)
A8 Training effects: 3.9 (1 = unimportant; 5 = very desirable)
A9 Statistical analysis: 3.8 (1 = unimportant; 5 = very desirable)
A10 Explanation function: 3.8 (1 = unimportant; 5 = very desirable)

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:

B1 Structured questionnaires:	3.8 (1 = inconvenient; 5 = very helpful)
B2 Input of findings:	3.7 (1 = insufficient; 5 = too differentiated)
B3 Reminder function:	3.8 (1 = unnecessary; 5 = very helpful)
B4 Use of help:	2.5 (1 = never; 5 = very often)
B5 Relevance of system diagnoses:	2.9 (1 = unimportant; 5 = important)
B6 Influence of system on own diagnoses:	2.2 (1 = unimportant; 5 = important)
B7 Standardization of nomenclature:	4.4 (1 = unimportant; 5 = important)
B8 Comparability of sonographic records:	4.5 (1 = unchanged; 5 = improved)

The answers to these questions show that expectations and experiences agree in many aspects: the standardization of nomenclature is most acknowledged by the examiners (A3 | B7, B8), the input procedure is well accepted (A2, A6 | B1, B2) and the reminder function of the program is perceived as helpful (A4 | B3). This is also true for the effect of the system diagnoses, which is perceived as not so important (A5 | B5, B6). A difference between expectations and experiences exists with respect to the explanation function, which was declared as rather desirable, but rarely used (A10 | B4). The expected training effect (A8) was compared with the experiences of 5 beginners and clearly confirmed the expectations. They all emphasised that the program's most positive effect was to conduct an examination in a complete and structured way as well as in a standardised and reasonable examination sequence. The diagnostic properties of the program had been of only medium or transitory interest during the learning phase.

Since the sonographic report is sent to the referring physicians in the clinic, we also asked their opinion about the significance of the new record types with a questionnaire that was answered by 35 clinicians:

C1 Differentiated report on organ findings:	3.8 (1 = never read; 5 = always read)
C2 System diagnoses:	3.0 (1 = never read; 5 = always read)
C3 Examiner comment:	4.9 (1 = unnecessary; 5 = necessary)
C4 Standardization of nomenclature:	3.5 (1 = unnecessary; 5 = necessary)
C5 New records:	3.6 (1 = neutral; 5 = positively acknowledged)
C6 Significance of sonographic records:	2.3 (1 = unchanged; 5 = enhanced)

The results indicate that the clinicians most strongly rely on the examiner comment, but often read the differentiated report with all observed findings and the automatically inferred system diagnoses. They welcome the new record and the standardization of nomenclature, but judge them only as a small enhancement.

3.2 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 or a higher quality of the reported diagnoses. To answer the first question, we entered the data of 103 hand written reports documented before the program's introduction into SC and noted whether all questions asked by SC could be answered with the available data. If not, two senior examiners judged the information gaps in the free text reports as relevant or dispensable. From the 287 information gaps found, the domain experts judged nearly half (132) as relevant. For the second question concerning the diagnoses, the senior examiners compared the diagnoses in the free text reports with those generated by the program and used their own diagnoses based on the same patient data as gold standard. They found 179 problematic diagnoses form a total of about 600 diagnoses (103 cases with 6 diagnoses on average) in the free text reports and 86 problematic diagnoses generated by SC.

Since this evaluation is difficult to interpret, because there is some scope to answer standardized questions from free text reports, we conducted a second evaluation, where we used 112 prospective, consecutive records and compared the documented conclusions of the examiners with those of SC. There were almost no relevant information gaps. This is due to the guided data acquisition strategy of SC, which had a significant clinical impact. The diagnostic conclusions were judged by three domain experts as "correct" or "problematic", when they all agreed on the same assessment. From the 412 diagnoses in these records (i.e. in this sample an average of 3.7 per case), the examiners missed 107 (26%) true diagnoses and stated an additional 32 diagnoses, which were not supported by the documented findings. In contrast, all diagnostic conclusions of SC were judged as adequate. When the 412 diagnoses are differentiated 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%) and stated an additional 15 unsupported diagnoses. Again the 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 high degree of problematic simple diagnoses (50 from 267, i.e. 19%) indicates documentation errors. Nevertheless the inconsistence between documented findings and diagnostic conclusions is rather high. This is quite astonishing, since the SC-diagnoses were visible to the examiners before writing their final comment. This fact is consistent with the low influence of the system's diagnoses on the own diagnoses of the examiners in the questionnaire (B6). To investigate this phenomenon further, we plan several steps: Follow-up evaluations to investigate the hypothesis that the acceptance of the system diagnoses by the examiners takes a longer period of time, an additional interface enabling the examiners to copy the system's diagnoses quite easily in their free text report and a critic component to compare the free text diagnoses and the system's diagnoses. The critic component generates warnings in case of serious discrepancies and offers generated forms for correction of the discrepancies. The latter requires an information extraction component for identifying coded diagnosis in a free text format.

3.3 Statistical Analysis

The physicians considered statistical analysis as one of the desirable features (A9) before the introduction of SC. Per month about 300 patient records are documented in detail. Typical questions concern correlations among pathological states of different organs (the intra organ relations are usually well known). The data mining technique of subgroup mining [Klösgen 02] is most suitable for questions like e.g. whether a certain pathological state is significantly more frequent if combinations of other pathological states exist. Since the efficient use of a subgroup data mining tool requires some experience and statistical knowledge, we decided for a two-step model for the clinical introduction. The first step consists of simple tool with standard reports similar to the OLAP interface to data warehouses [Han & Kamber 00] with the option of manual refinement of the data along predefined hierarchies (e.g. time hierarchies like day, week, month; general patient attributes like age, gender and body weight with categories and diagnostic hierarchies like organs and special diseases concerning an organ). This simple interface is directly integrated into the GUI of SonoConsult. The second step consists of a powerful tool VIKAMINE designed 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), the availability of much background knowledge (which should not be rediscovered but used to find new, often subtle correlations) and the importance of controlling confounding factors (like age, gender, body weight etc.). It offers different search options for automatic subgroup discovery and various interactive visualizations for active user involvement. Both tools operate on the same data base. When the user detects something unexpected in the data with the simple tool, he or she may switch to VIKAMINE to analyse these unexpected features in detail, e.g. to discover subgroups corresponding to the unexpected features which might serve as an explanation. First experiments with VIKAMINE were quite promising [Atzmüller et al. 04], but the tool is currently not in routine use.

4 Summary and Further Work

The evaluations of SC 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 analysis. Although the system was well accepted in general, its diagnostic conclusions were largely ignored. This evaluation result requires further investigations.

References

Atzmüller, M., Puppe, F., Buscher, H.-P.: Towards Knowledge-Intensive Subgroup Discovery, LWA-Workshop 2004 (Lernen, Wissensentdeckung und Adaptivität), Berlin 2004

Berner, E., Maisiak, R., Cobbs, G., Taunton, O.: Effects of a Decision Support System on Physicians´ Diagnostic Performance. J Am Med Inform Assoc 6, 420-427, 1999.

Buscher, H. P., Engler, C., Fuhrer, A., Kirschke, S., Puppe, F.: HepatoConsult: a Knowledge-Based Second Opinion and Documentation System. Artif. Intell. Med. 24, 205-216, 2002.

Darmoni S., Poynard T.: Computer-Aided Decision Support in Hepatology. Scand J Gastroenterol 27, 889-896, 1992.

Han, J., and Kamber, M.: Data Mining, Concepts and Techniques, Chap. 2, 2000.

Huettig M, Buscher G, Menzel T, Scheppach W, Puppe F., Buscher, H.P.: A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography, Med Klin 99: 117-122, 2004.

Klösgen, W: Handbook of Data Mining and Knowledge Discovery, Chapter 16.3 Subgroup Discovery, Oxford University Press, 2002.

McDonald C.: Medical Heuristics: The Silent Adjudicators of Clinical Practice. Ann. Intern. Med. 124, 56-62, 1996.

Puppe, F.: Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-KIT D3. Int Journal of Human-Computer Studies 49, 627-649, 1998.