TUCS

Helena Karsten | Barbro Back
Tapio Salakoski |Sanna Salanterä
Hanna Suominen (Eds.)

Proceedings of

Louhi'08

The First Conference on Text and Data Mining of Clinical Documents

Turku Centre for Computer Science

TUCS General Publication
No 52, September 2008

TUCS

Proceedings of

Louhi’08

The First Conference on Text and Data Mining of Clinical Documents

September 3rd - 4th, 2008, Turku, Finland

Editors:

Helena Karsten
Barbro Back
Tapio Salakoski
Sanna Salanterä
Hanna Suominen

TUCS General Publication
No 52, September 2008
Digital Edition

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Foreword

The First Louhi Conference on Text and Data Mining of Clinical Documents (Louhi'08) seeks to bring together researchers and practitioners into a multi-disciplinary conference on new application areas for intelligent systems. The often repeated challenges to health care include also the vast amounts of documentation produced during care and as outcomes of health and medical research. For the practitioners to manage with these amounts, many intermediaries – such as summaries, code lists and rulebooks - are in use. To turn health and medical research usable also to practitioners, medical journals present concise reports, medical databases provide search engines and classifications for finding relevant information, and evidence based guidelines form the rules to follow. The methods and algorithms presented in this conference seek to give computational tools for improving the automated processes and for supporting the manual work in forming these intermediaries.

What make the Louhi Conference and the Louhi project[1] unique is that we seek to constantly reflect our work onto the actual work carried out in hospitals and clinics (Suominen et al 2005). Thus, daily and discharge summaries in a hospital ward are more than just summaries; they are carefully constructed assessments of care given so far (Berg 2004). Code lists such as ICD-10 not only enumerate illnesses, they also tell of the culture of care giving (Bowker & Star 1999). Clinical guidelines not only provide state of the art advice on certain situations and conditions, but following them may also have unintended consequences (Timmermans & Berg 2003). We are only beginning to find out the ethical quagmires of using intelligent systems in actual care (Suominen et al 2007). We have also learned that the work practices in the busy hospitals are quite fragile, especially when technologies around them break down (Forsell, Karsten & Vuokko 2007; Karsten & Vuokko 2008).

The Proceedings of this very first international Louhi conference show some of the multitude and complexity of the issues we are facing. In the five full research papers we can read many of the challenges outlined. In the two first papers, evidence-based guidelines are to be constructed with the help of research databases and analyses of patient records. The third paper seeks to build a syntactic parse to enable consequent semantic analysis of patient records. The fourth paper deals with improving the quality of information in diagnosis-based registries. The fifth paper moves the discussion to a hospital by presenting four actual cases of technology use in wards and by outlining lessons for future technology implementations, including intelligent systems.

[1] http://www.med.utu.fi/hoitotiede/tutkimus/tutkimusprojektit/louhi/

In this kind of a multi-disciplinary, emerging field, it is natural that most of the papers and hence presentations are of work-in-progress. To contrast medical entries with nursing narratives, possibly in large quantities, provides innovative approaches to improve practice. The second short paper describes an automatic encoding system. The coding schemes, even though useful for readers of patient records, can be tedious to use manually. In medical emergencies, chief complaints outline the subsequent diagnostic work. A case-based re-writing system for these is a step towards computational understanding and processing of the cases. The final short paper present a method for dividing nursing narratives into non-overlapping segments which then can be labelled and sorted for a better overview of the patient, for example for writing a summary of care. Two Work-in-Progress Proposals conclude the proceedings. Computer-based decision support is analysed with a practical case. A research proposal is presented for pre-processing clinical narratives.

Dear reader, you are welcome to peruse these proceedings. We hope they will open new vistas and evoke new research problems to tackle.

August 2008

References

Berg, M., Ed. (2004). Health Information Management. Integrating Information Technology in Health Care Work. London, Routledge.

Bowker, G. and S. L. Star (1999). Sorting things out: classification and its consequences. Cambridge, MA, USA, MIT Press.

Suominen HJ, Lehtikunnas T, Hiissa M, Back B, Karsten H, Salakoski T, Salanterä S. *Natural language processing for nursing documentation*. In: Fonseca JM, editor. Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare; 2005 June 29 - July 1: Costa da Caparica, Portugal; 2005. p. 147-54.

Suominen H, Lehtikunnas T, Back B, Karsten H, Salakoski T, Salanterä S. *Applying language technology to nursing documents: pros and cons with a focus on ethics*. International Journal of Medical Informatics, 2007 Vol 76S2, p. S293-S301.

Timmermans, S. and M. Berg (2003). The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care. Philadelphia, PA, USA, Temple University Press.

Turku, Finland, August 2008
Helena Karsten

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Conference chair

Sanna Salanterä, University of Turku

Program co-chairs

Barbro Back, TUCS and Åbo Akademi University
Tapio Salakoski, TUCS and University of Turku

Publications chair

Helena Karsten, TUCS and Åbo Akademi University

Programme committee

Alexey Tsymbal, Siemens AG, Germany
Andrew B Clegg, National Collaborating Centre for Women's and Children's Health, UK
Anne Scott, Dublin City University, Ireland
Anneli Ensio, University of Kuopio, Finland
Anthony Maeder, Australian e-Health Research Centre/CSIRO ICT Centre, Australia
Antonina Durfee, RBS Citizens, N.A., Rhode Island, USA
Filip Ginter, University of Turku, Finland
Helena Karsten, TUCS and Åbo Akademi, Finland
Jari Forsström, Tampere University Hospital, Finland
Kaija Saranto, University of Kuopio, Finland
Liisa von Hellens, Griffith University, Australia
Minna Kaila, University of Tampere, Finland
Mykola Pechenizkiy, Eindhoven University of Technology, Netherlands
Richárd Farkas, University of Szeged, Hungary
Sanna Salanterä, University of Turku, Finland
Seppo Puuronen, University of Jyväskylä, Finland
Shuhua Liu, Stanford University, USA
Simo Vihjanen, Lingsoft, Finland
Tapio Pahikkala, TUCS and University of Turku, Finland
Tomas Eklund, Griffith University, Australia
Walter Sermeus, Catholic University Leuven, Belgium
William TF Goossen, Results 4 Care, Netherlands
Veronika Laippala, University of Turku, Finland
Vladimir Estivill-Castro, Griffith University, Australia

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Invited speakers

Filip Ginter, University of Turku, Finland
Kaarina Tanttu, Hospital District of Southwest Finland, Turku University Hospital, Finland
Richárd Farkas, University of Szeged, Hungary

Organizers

Louhi project: http://www.med.utu.fi/hoitotiede/tutkimus/tutkimusprojektit/louhi/
Turku Centre for Computer Science (TUCS)
University of Turku
    Department of Information Technology
    Department of Nursing Science
Åbo Akademi University
    Department of Information Technologies

Local organizing chair

Hanna Suominen, TUCS and University of Turku

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Sponsors

City of Turku

Lingsoft Inc.

Turku Centre for Computer Science

Turku University Foundation

Åbo Akademi University
Department of Information Technologies

Federation of Finnish Learned Societies

Stiftelsen för Åbo Akademi

Turku Science Park

University of Turku
Department of Information Technology
Department of Nursing Science

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Contents

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Research Papers

Research papers are double-blind peer-reviewed and they are presented by the authors in a regular conference session.

Streamlining the Clinical Guideline Production Process with Fuzzy Citation Matching

Andrew B. Clegg[1] and Debbie Pledge[2]

[1] Research Dept. of Structural and Molecular Biology, University College London,
Gower St., London, WC1E 6BT
andrew.clegg@gmail.com
[2] National Collaborating Centre for Women's and Children's Health,
4th Floor, King's Court, 2-16 Goodge St, London, W1T 2QA
dpledge@ncc-wch.org.uk

Abstract. We present a novel application of text mining in clinical practice guideline development. In order to evaluate the benefits for the National Collaborating Centre for Women's and Children's Health of subscribing to the PsycINFO literature database, we needed to determine how much relevant and eligible evidence PsycINFO provides the Centre's systematic reviewers with, over and above the material already available through MEDLINE, EMBASE etc. The lack of inter-database identifiers led us to develop a simple, fast and accurate citation matching algorithm capable of precision and recall scores simultaneously higher than 90%. Applying this algorithm retrospectively to the literature searches performed for three published guideline projects, we found that only 6 papers of sufficient quality and relevance for systematic reviewing (0.3%) had been found uniquely via PsycINFO. This data, acquired by the citation matching algorithm, suggests that, for the NCC-WCH's purposes, PsycINFO gives poor value for its time and money costs.

All source code written for this project is available from the authors on request.

1 Introduction

The development of evidence-based guidelines is an area that text and language technologies have yet to make much impact in, despite the fact that this activity relies on highly manual and knowledge-intensive equivalents of information retrieval, information extraction and text mining. While there is a natural reluctance in the medical domain to leave clinically-important data in the hands of inaccurate, unsupervised machines, there are also problems of data integration and knowledge management that can only be solved through text mining techniques. We present the results of a practical investigation that helped us identify time and cost wastage in the guideline development process. This relied on a novel application of a simple text processing algorithm to the problem of joining records between heterogeneous literature databases. We also describe a means to obtain training and test data for such techniques which does not require manual annotation.

1.1 Background

As the principles of evidence-based medicine (EBM) become more and more widespread in clinical practice, healthcare providers are increasingly relying on the work of specialist research groups to track, analyse and interpret the literature. One such body, the National Collaborating Centre for Women's and Children's Health[3] (NCC-WCH), is commissioned by the National Institute for Health and Clinical Excellence[4] (NICE) to produce clinical guidelines for the National Health Service[5] (NHS) in England and Wales. In order for the findings of primary research to benefit patients, all the available evidence on a given topic must be critically appraised and systematically reviewed, so that experimental results are not presented in isolation but taken as a coherent body of evidence which can guide clinical decision-making. In this context, secondary medical research has grown into a scientific discipline in its own right, with an array of statistical and methodological techniques for assessing study quality, detecting bias and aggregating quantitative data from different sources.

Of course, the raw evidence which underlies such analyses and syntheses has to come from somewhere, and in most healthcare contexts it is to be found primarily in peer-reviewed publications such as journal articles, conference proceedings and existing systematic reviews. Locating and acquiring this evidence in a transparent and unbiased manner requires considerable methodological rigour in itself, as well as access to comprehensive information resources. While it is true of all the sciences that the published literature both supports and steers new work, missing a key research finding in the production of a systematic review or clinical practice guideline could have far-reaching consequences that are both dangerous and costly. This is especially true of state-funded guidelines (such as NICE's) that are rolled out across regions or nations, or used to shape public health policy at the governmental level.

MEDLINE[6] often seems to be regarded as a universal point of access to the published literature in the life sciences, due in part no doubt to its free availability via PubMed[7] and other services. However, informatics scientists working in this field need to look much further afield to achieve comprehensive coverage. There are many other literature databases which might hold crucial information unavailable in MEDLINE, for a variety of reasons. Some resources index more publications in specialist fields (e.g. health economics or psychology), non-journal formats (e.g. book chapters or theses), or foreign languages. Others allow the publication of experimental protocols for studies in progress, for example, or provide value-added features such as structured abstracts or mini-reviews of new papers. While most of these resources are small compared to MEDLINE,

[3] `http://www.ncc-wch.org.uk/`

[4] `http://www.nice.org.uk/`

[5] `http://www.nhs.uk/`

[6] `http://www.nlm.nih.gov/pubs/factsheets/medline.html`

[7] `http://www.pubmed.org/`

some, such as EMBASE,[8] are of a similar order of magnitude. In general, there are no figures available which quantify the overlap between these resources.

Naturally, the very diversity of such databases is a constant issue for information scientists and systematic reviewers. Although there are *de facto* standards, almost every database has its own indexing policy, controlled (or uncontrolled) vocabulary, file format, logical operator syntax, field names and idiosyncratic interface. The quality and consistency of indexing and even data entry can vary considerably within one database and even more so between them. Although third-party search engines like Ovid[9] and bibliographic tools like Reference Manager[10] provide some assistance, they introduce issues and costs of their own, and sometimes risk obscuring important differences rather than genuinely achieving interoperability. For example, Reference Manager has a function to filter out duplicate citations automatically, but this does not work reliably when the records are from different databases. And although Ovid provides a superficially consistent query interface across most of the databases it indexes, this illusion of consistency in fact conceals from the user the fact that seemingly-identical queries have different semantics in different databases.

1.2 Time vs Cost vs Information

Unlike groups such as the Cochrane Collaboration[11] who produce systematic reviews of the evidence on a single clinical scenario at a time (e.g. "Allopurinol for preventing mortality and morbidity in newborn infants with suspected hypoxic-ischaemic encephalopathy" [1]), the NCC-WCH produces much larger guidelines on broader topics. These use systematic literature reviews, backed up with statistical meta-analysis and economic modelling where necessary, to derive best-practice recommendations for healthcare professionals and information for their patients. Criteria for eligibility of evidence, based on applicability and quality, are formalized in advance to reduce the risk of bias. Each guideline addresses a number of distinct clinical questions, each of which necessitates the development of a search strategy for every database likely to contain evidence. An example of such a strategy is shown in Table 1 and illustrates the scale and complexity of the queries involved.

The NCC-WCH follows a methodology set out in the NICE guidelines manual.[12] Core databases—MEDLINE, EMBASE, CINAHL[13] and the Cochrane Library[14]—must be searched for every clinical question. Subject-specific databases—of which PsycINFO[15] is one—are searched when they are deemed to be relevant to the question. PsycINFO (see Sect. 1.3) is most likely to be searched when

[8] `http://www.elsevier.com/wps/product/cws_home/523328`

[9] `http://www.ovid.com/`

[10] `http://www.refman.com/`

[11] `http://www.cochrane.org/`

[12] `http://www.nice.org.uk/guidelinesmanual`

[13] `http://www.cinahl.com/`

[14] `http://www.thecochranelibrary.com/`

[15] `http://www.apa.org/psycinfo/`

the question covers the topics of communication, support, information, patient's views and experiences, or psychological/behavioural interventions.

Table 1. An extract from a search strategy for a project about reducing social inequalities in immunization uptake, designed for MEDLINE via Ovid. The complete 35-line strategy is in fact fairly short by the NCC-WCH's standards.

1. ((improv$ or enhanc$ or encourag$ or increas$ or support or assist$ or maximi$ or promot$) adj2 (uptake or coverage or cover or acceptanc$ or complianc$ or adoption or rate$ or access or equit$ or equalit$)).tw.
2. ((decreas$ or reduc$ or minimi$ or detect$ or chang$) adj2 (inequalit$ or unequal or inequit$ or disparit$ or imparit$ or varia$ or unevenness or discrepanc$ or imbalance$ or differen$ or barrier$)).tw.
3. ((geographic or regional or urban or rural or ethnic or religio$ or class or socioeconomic or social or economic or demographic or cultural) adj2 (inequalit$ or unequal or inequit$ or disparit$ or imparit$ or varia$ or unevenness or discrepanc$ or imbalance$ or differen$ or barrier$)).tw.
4. UNEMPLOYMENT/ or exp SOCIOECONOMIC FACTORS/ or URBAN POPULATION/ or SUBURBAN POPULATION/ or RURAL POPULATION/ or VULNERABLE POPULATIONS/ or SINGLE PARENT/ or DISABLED CHILDREN/
5. (poor or income or unemploy$ or middle class or working class or disadvantaged or socially excluded or inner city or poverty or deprived or vulnerable or jobless or immigrant$ or migrant$ or asylum seeker$ or refugee$ or single parent$ or single mother$ or lone parent$ or lone mother$ or disabled or disabilit$ or handicap$).tw.
6. PUBLIC HOUSING/ or exp SOCIAL WELFARE/ or exp SOCIAL WORK/
7. (social housing or council hous$ or council estate or temporary accommodation or foster home or orphanage or looked-after child$ or housing benefit or income support or unemployment benefit or jobseeker's allowance or child benefit or social services or social work$ or social security).tw.
8. MINORITY GROUPS/ or exp RELIGION/ or exp ETHNIC GROUPS/ or exp CONTINENTAL POPULATION GROUPS/
9. (minorit$ or ethnic$ or black$ or white$ or asian$ or indian$ or pakistani$ or bangladeshi$ or african$ or caribbean$ or chinese or oriental$ or turk$ or african$ or arab$ or gyps$ or romany or travel?er$ or religio$ or christian$ or catholic$ or protestant$ or jew$ or muslim$ or hindu$ or sikh$ or buddhist$ or jehovah's witness$).tw.
10. or/1-9
11. exp IMMUNIZATION/ or exp IMMUNIZATION PROGRAMS/
12. (immuni$ not innate immunity).tw.
13. (vaccin$ not vaccinia).tw.
14. or/11-13
15. 10 and 14
16. exp TRAVEL/
17. 15 not 16
18. limit 17 to (english language and humans)

As a result of the heterogeneity issues described earlier, adapting a search strategy from one database to another is time-consuming and increases the chances of human error. Importantly, the time taken to adapt a strategy is not dependent on the quantity or quality of the information that it will yield. Obviously, these cannot be known in advance, and nor can the degree of redundancy between two equivalent queries in different databases. However, all three factors can be investigated retrospectively, and given that time and funding are finite and must be spent wisely, this approach suggests a means to assess the value of including a particular database in a systematic literature search. We developed this idea in order to assess how worthwhile the recent subscription to the PsycINFO database had been.

Our analysis set out to answer the following key questions:

1. How many references did PsycINFO yield that were not returned by the equivalent searches on the existing repertoire of databases?
2. Of these, how many were judged relevant by a systematic reviewer, based on title and abstract?
3. Of these, how many then met the criteria for applicability and quality based on the full text of the paper, and were thus used as evidence?

Answering these questions would be trivial if not for one problem. There is no global identifier that allows a record in one database to be uniquely and unambiguously located in another. The results for each search are downloaded separately before pooling, but without a noise-tolerant matching method, a manual inspection would have been necessary for each search in order just to get past question 1. Since several of the searches under examination had yielded over 1,000 hits across all databases, this would not have been feasible in the time available.

1.3 The PsycINFO Database

The American Psychological Association's PsycINFO database contains, at the time of writing, more than 2.5 million bibliographic records for publications from 1806 to the present day. 78% of these are journal articles, from journals dedicated to psychology, psychiatry and related topics, as well as selected relevant articles from other periodicals in fields as diverse as paediatrics and artificial intelligence. Almost all articles are peer-reviewed, and although most are in English, articles in other languages are covered if the titles, abstracts and keywords are provided in English. The other 22% of entries are a roughly even split between edited books and chapters, and secondary publications such as dissertations. The vast majority of entries have abstracts, many have full citation lists and/or tables of contents, and all have metadata supplied by the APA's indexers, such as publication type, controlled-vocabulary keywords, geographical tags, experimental population information, and sponsorship details, as appropriate. The NCC-WCH's subscription to PsycINFO, via Ovid, costs £1857 annually at the time of writing.

It is clear then that PsycINFO differs from MEDLINE, for example, not just in topic focus but also in publication type (MEDLINE covers journals only), date range (the oldest MEDLINE records are from 1949), and kinds of metadata included. However, most of the metadata is ignored in NICE's systematic reviewing methodology; in general only title, abstract and date, and sometimes keywords, are used to determine relevance before the full text of the paper is retrieved. Furthermore, although a significant number of the clinical questions we seek to answer are psychological in nature or have a psychological component, it is not clear a priori that the additional coverage provided by PsycINFO actually yields further useful material beyond that which is available via the core databases mandated by NICE. We embarked upon the investigation reported in this paper in order to clarify this issue empirically.

1.4 Related Work

Although there have been a number of experiments to compare the content of medical literature databases, we are not aware of any with similar scope, intention or methods to this one. Most concentrate on the two largest resources, MEDLINE and EMBASE, and use manual methods which do not scale well. Wilkins et al. [2] compared the results of several complex searches in these databases using the EndNote bibliography application, finding a low degree of overlap and a larger number of results from EMBASE. However, their matching method is not described in detail; no evaluation of accuracy was performed and it is unclear what level of manual inspection was necessary. Suarez-Almazor et al. [3] also compared coverage in MEDLINE and EMBASE, using a method relying on hand-searching. A different kind of comparison of the same two databases was carried out by Sampson et al. [4] who looked at the statistical effects of including studies from either one or both of the databases when pooling results for a meta-analysis. They conclude that papers indexed only in EMBASE make a small but significant conclusion to the overall results of meta-analyses. Royle and Milne [5] compared several databases for their coverage of randomized controlled trials used in Cochrane reviews, by manual searching which involved searching on author names and title keywords and inspecting the results. They found, perhaps unsurprisingly, that Cochrane's own register of controlled trials was the best source of such studies. McDonald et al. [6] looked at several databases' coverage of psychiatry articles, but only by comparing lists of journals covered by each database rather than specific citations received in response to an actual search.

There is also a sizable body of literature on citation matching, much of which uses methods rather more complicated than those presented here, based on either probabilistic models (e.g. Pasula et al. [7], Wellner et al. [8]) or deterministic algorithms (e.g. Lawrence et al. [9], Monge and Elkan [10]). However, our investigation is not directly comparable to these projects since the citations we are interested in are already segmented into fields. Most of the research work on citation matching concentrates on the harder problem of connecting free-text citations extracted from the reference lists of papers to the actual articles they refer to, without explicit field markers or labels. This necessitates the extra levels of complexity found within such work. The fact that the data for our investigation was already in the form of structured records reduced the problem to a much simpler task. This enabled us to rapidly build a solution good enough to guide business decisions within the NCC-WCH that requires no machine learning and little calibration, and can be implemented in 13 kilobytes of Perl.

2 Methods

This section describes the fuzzy citation matching algorithm we developed; the steps we took to acquire test data for it automatically; the calibration process required to tune its single parameter; and finally, the protocol for tracing papers from PsycINFO through the reviewing workflow.

2.1 Fuzzy Citation Matching

The task of matching citations to the same paper in different databases is problematic for several reasons, including:

- Any two records from different literature databases will not, in general, have the same fields.
- Those fields that both have in common are not guaranteed to both be filled in.
- Those equivalent fields that are filled in for each record are not guaranteed to use the same format. See Table 2.
- Import/export and file format conversion processes can corrupt or truncate fields. We frequently find author lists that have been truncated after the first or second author, without "et al." as a placeholder.
- Indexers working for database publishers often insert their own notes or metadata into existing fields, even though the definitions of these fields do not logically include the inserted strings. Examples include "[Spanish]" or "[48 refs.]" appended to a title, or "(PsycINFO Database Record (c) 2003 APA, all rights reserved)" appended to an abstract. Since these strings are not in fact part of the title or abstract, it would be more sensible to put them into entirely separate fields, but many databases nonetheless follow this practice.
- Staff at some databases write, rewrite or summarize abstracts themselves, rather than using those supplied by the publishers. For example, in the MEDLINE entry for the article in Table 2, the abstract is 309 words long, but the PsycINFO abstract has only 193 words.
- Although universal document/resource identifiers do exist (for example Digital Object Identifiers[16]), they are not in widespread use. Of all the databases covered in this study, only EMBASE includes DOI codes and then only when available.

[16] http://www.doi.org/

Table 2. A comparison of some equivalent fields taken verbatim from the MEDLINE (left) and PsycINFO (right) entries for the same paper [11], showing variation in content and formatting.

Goode PS. Burgio KL. Locher JL. Roth DL. Umlauf MG. Richter HE. Varner RE. Lloyd LK.	Goode, Patricia S; Burgio, Kathryn L; Locher, Julie L; Roth, David L; Umlauf, Mary G; Richter, Holly E; Varner, REdward; Lloyd, LKeith.
Effect of behavioral training with or without pelvic floor electrical stimulation on stress incontinence in women: a randomized controlled trial.[see comment].	Effect of behavioral training with or without pelvic floor electrical stimulation on stress incontinence in women: A randomized controlled trial. [References].
JAMA. 290(3):345-52, 2003 Jul 16.	JAMA: Journal of the American Medical Association. Vol 290(3) Jul 2003, 345-352. American Medical Assn, US
Clinical Trial. Journal Article. Randomized Controlled Trial.	Empirical Study; Clinical Trial; Journal Article

The solution we arrived at is deterministic, fast, easy to implement, and requires no hand-annotated training data and little parameter tuning. It is based on the standard Levenshtein edit distance, which is a measure of the difference of two strings. This is defined as the smallest number of character insertions, deletions or substitutions required to turn one string into the other. Our implementation requires a tolerance threshold parameter or ‘fuzz factor’ d, which defines how much variation is allowed between two strings for them still to count as a match. d is interpreted as a percentage of the total length of the two strings, so the amount of variation allowed scales with the amount of text to match. For example, if we were to compare two strings of length 11 and 9 with $d = 10$, they would count as a match only if one could be turned into the other within two edits (10% of total string length 20).

Given two database records to match, our algorithm proceeds as follows:

1. Extract the title field from each record. If either is blank, return MISMATCH, otherwise. . .
2. Normalize the titles (see below).
3. Calculate the Levenshtein distance between the normalized titles. If they are more different than d allows, return MISMATCH, otherwise. . .
4. Extract only the *first* author’s surname from each record. For simplicity here we just use the first complete string of non-whitespace characters.
5. Normalize the surnames (as before, see below).
6. Calculate the Levenshtein distance between the normalized surnames. If they are more different than d allows, return MISMATCH, otherwise return MATCH.

One point to note is that if one or both records has no authors, the match is allowed to proceed as normal, with empty author fields just counting as strings of length 0. However a missing title on either side causes an immediate fail. This is because papers are much more likely to share a first author name by chance than an entire title. In practice it seems unlikely that bibliographic records with missing titles would slip unnoticed into a production system, since titles are used so extensively by humans to identify documents.

An important part of our algorithm is the string normalization function, which accounts for many of the more predictable variations between strings (particularly titles), meaning that d does not need to be set as high as it would have been if operating on unprocessed database fields. The normalization function proceeds as follows:

1. Convert entire string to lowercase.
2. Remove any character (including whitespace) that isn't a letter or a bracket: () []
3. If the string starts with an open-square-bracket, find its partner and remove them both. Foreign-language titles translated into English are often square-bracketed.
4. If the string ends with a bracketed sequence of characters, remove it entirely. Repeat until there are no more bracketed sequences at the end. This accounts for metadata tagged onto the end of title fields such as "(Abstract)" or "[Review] [28 refs]".
5. Remove any remaining bracket characters.

We will not describe the dynamic programming algorithm for computing the Levenshtein distance here as it is well covered by textbooks (e.g. [12]). Indeed, implementations are either supplied as standard or available to install for most programming languages. Our Perl script uses the String::Levenshtein module[17] which makes use of native C code internally for speed.

2.2 Acquiring Test Data

Although there is no ID or accession number which is shared across all the databases, CINAHL does include the PubMed ID (PMID) for some articles, enabling these to be linked to the matching records in MEDLINE. This provided us with a means to generate large quantities of test data in a mostly automatic manner, with no manual matching required. Using Ovid, we ran ten queries in parallel against MEDLINE and CINAHL, and downloaded the 1000 most recent results for each query in each database. Nine of the queries were on topics that the NCC-WCH has worked on guidelines for, spanning a range of issues in clinical medicine and public health. The tenth simply covers the last 1000 papers in 2007 (any subject, any journal). We verified that every record in each MEDLINE file had a PMID and that there were no duplicates in any MEDLINE file. For CINAHL, the number of records with a PMID varied between topics, and there was a small number of duplicate records (see Table 3). The set of unique

[17] `http://world.std.com/~swmcd/steven/perl/lib/String/Levenshtein/Levenshtein.html`

PMIDs common to both MEDLINE and CINAHL for each topic represents the population of genuine, verifiable matches that we tested the algorithm on its ability to identify.

Table 3. 10 test sets of partially-overlapping datasets from MEDLINE and CINAHL. Each set contains 1000 results from each of these two databases.

Name	PMIDs in CINAHL	Unique PMIDs in CINAHL	Unique PMIDs common to CINAHL & MEDLINE
constipation	534	533	280
diarrhea & vomiting	564	564	162
feverish illness	606	606	152
immunization	699	699	147
labour	517	517	314
meningitis	659	658	107
personal, sexual & health ed.	576	575	309
urinary incontinence	472	472	182
urinary tract infection	561	561	132
2007, last 1000 papers	302	302	13

2.3 Calibrating the Algorithm

For each of the ten topics, we extracted the common set C of PMIDs that were present in both the MEDLINE file and the CINAHL file. Our test script then ran the matching algorithm over the two files 21 times, with the fuzz factor ranging from 0 to 100 in increments of 5. For each run, the script extracted all reported matches involving at least one record with a PMID found in C. Of these, any matches between two records with the same PMID were counted as true positives, and any matches from a MEDLINE record to a CINAHL record with a different or missing PMID were counted as false positives. The script ignored all matches where neither record had a PMID found in C. The false negatives were calculated by subtracting the number of true positives from the size of C.

Using the counts of true positives, false positives and false negatives for each run, our script then calculated precision (P, the proportion of reported matches which were correct) and recall (R, the proportion of real matches which were reported). Plotting these 21 precision-recall pairs for each of the ten topics gave us ten curves which showed the accuracy of the algorithm over a range of possible fuzz factors (see Results & Discussion). This enabled us to select an appropriate fuzz factor for our main investigation.

2.4 Analysing PsycINFO

In order to assess the contribution of PsycINFO to the guideline development process, we retrieved from the NCC-WCH's archives the search records for 21 clinical questions from previously-published guidelines (Table 4) where PsycINFO had been searched in addition to the existing core databases. The guideline topics were as follows: *HMB* = heavy menstrual bleeding [13], *ipc* = intrapartum care [14], and *UI* = urinary incontinence [15]. For each of these search result sets, we set out to supply each of the following pieces of information:

1. The number of hits retrieved by the PsycINFO search strategy.
2. The proportion of records in 1 that were unique to PsycINFO.
3. The proportion of records in 2 that were judged relevant after manual inspection by a systematic reviewer.
4. The proportion of records in 3 referring to an article (of whatever kind) that was ultimately included in a systematic review for a clinical question.

Thankfully a careful audit trail was kept of the search and retrieval process for each question, enabling these values to be discovered with the help of the citation matching algorithm. Indeed, the value for step 1 above could be retrieved without the use of the algorithm, simply by counting the number of PsycINFO records downloaded from Ovid at the time, and these values are given in Table 4. It can be seen that there is a great deal of variation between these data sets, not just in the overall numbers of results and the numbers of results found in PsycINFO, but also in the proportion of the overall results which come from PsycINFO. This is as low as 1% for many questions but as high as 23% for the *alt* topic (complementary and alternative therapies). Intuitively, this is to be expected, since different health topics which have psychological aspects ought to vary in the extent to which their study is tractable and interesting to the psychology community.

The answer to step 2, the proportion of *unique* PsycINFO hits, could be determined by citation-matching the PsycINFO results for each search against the set of all records downloaded from all of the other databases, and counting the number of unmatched PsycINFO records. For step 3, the proportion of *relevant* hits unique to PsycINFO, we citation-matched the unique PsycINFO records from the previous step for each topic against the database of citations retained by a reviewer after manual relevance filtering (known as ‘weeding’).

As well as the search and weeding audit trails, an Access database is kept for every guideline which identifies the full-text articles ordered for systematic reviewing (i.e. those which made it through the weeding process) and specifies for each article whether it was included in the review or excluded for methodological reasons. The reasons for excluding a paper vary from question to question and guideline to guideline, but usually fall into several categories. One is inapplicability to the target population of the review, for example mismatches in age range or nutrition, presence of comorbidities, or substantial differences in disease progression. Another is incompatibilities between the interventions (therapies) or outcomes (indicators of success) tested in the study and those which the review is concerned with. Serious study design or implementation issues, e.g. failure to report or follow experimental procedure properly, or account for potential sources of bias, can also lead to rejection. Also, the reviewers will

Table 4. The data sets (Ovid search results) from previously-published clinical guidelines that were used in our investigation into PsycINFO. The 'deduped' column shows the number of articles in each data set after applying Reference Manager's rather insensitive duplicate-removal algorithm.

Guideline	Clinical question	Raw hits in all databases	Deduped hits in all databases	Hits in PsycINFO
HMB	*ed*	170	123	22
HMB	*pe*	1955	1625	317
HMB	*peh*	541	511	39
HMB	*qol*	452	317	22
HMB	*rf*	780	618	42
HMB	*rfac*	381	315	2
IPC	*cc*	537	360	5
IPC	*comm*	3211	2584	291
IPC	*hvb*	279	186	3
IPC	*int1*	3288	1865	16
IPC	*int2*	1192	699	15
IPC	*pain*	3545	2126	141
IPC	*pb*	412	265	5
IPC	*supp*	817	546	30
UI	*alt*	173	150	39
UI	*cont*	2957	2365	15
UI	*life*	1343	1135	207
UI	*pbt*	1283	794	40
UI	*pharm*	2035	1315	12
UI	*ques*	586	413	11
UI	*tens*	77	57	1

sometimes exclude a paper if a particular question can be answered satisfactorily by other studies with a superior design, by reference to a standardized hierarchy of evidence set out in the NICE methodology manual which ranks study types according to their susceptibility to subjectivity and bias. For example, it may not be necessary to systematically review observational studies of a phenomenon or intervention that has also been investigated by randomized controlled trials. Finally, if the NCC-WCH's systematic review includes an existing review as evidence, those papers which have been included in the existing review will be excluded from the new one, in order to avoid their data counting twice.

From these databases we extracted bibliographic information for all of the articles which had been included in a systematic review, and once again used the citation matcher to identify those papers unique to PsycINFO by comparison with the results from step 3. Note that we did not restrict this search to the specific systematic review each paper was originally obtained for. Occasionally a reviewer identifies an article as providing evidence for a clinical question other than the one for which it was originally obtained, and PsycINFO should be given as much credit for providing such serendipitous evidence as it would be for any other evidence unobtainable elsewhere. The results of this analysis are presented in the next section.

3 Results & Discussion

This section presents the information obtained in the calibration process, discusses what these findings implied for the PsycINFO investigation, and finally describes the results of that investigation.

3.1 Calibration

After obtaining the precision-recall points for each of the ten test topics, we plotted all the resulting curves on a graph, along with a curve taking the mean precision and recall across all topics for each fuzz factor value (Fig. 1). It can be seen straight away that the algorithm proved capable of achieving P and R of over 90% simultaneously on most of the runs, with its single best score point (by F-measure, see below) being $P = 99.7\%$, $R = 93.2\%$ for a fuzz factor of 10 on the *personal, sexual & health ed.* topic. The two lowest curves represent the *feverish illness* and *immunization* topics. It is not clear why performance on these two topics was slightly poorer than the rest, but since neither topic was covered by any of the three guidelines in our main PsycINFO investigation, we did not perform a detailed error analysis. The *2007* topic was unusual, since in this case the two 1000-record sets from MEDLINE and CINAHL had only 13 PMIDs in common—in the other topics, this figure ranged from 107 (*meningitis*) to 314 (*labour*). As a result, the corresponding curve on the graph (a horizontal line at $P = 92.3\%$) should be seen as a rather pathological case.

3.2 Selecting an Optimum Fuzz Factor

Although the calibration results demonstrated that the matching algorithm is accurate enough to support an experiment like this, they did leave the question of where exactly to set the fuzziness threshold somewhat open. The mean P-R curve shows a gentle transition from favouring one score to the other, rather than a sudden swing, and we had no reason for deliberately optimizing for precision or recall that was justifiable in the context of our experimental goals. Choosing a conservative matching strategy (higher precision), for example, would reduce the risk of overestimating the overlap between PsycINFO records and others, but it would simultaneously increase the risk of failing to identify that a given record had come from PsycINFO in the first place.

In order to make our choice of cutoff point more methodical, then, we calculated the balanced F-measure for each of the 21 mean precision-recall points. This is the harmonic mean of precision and recall, which behaves as a weighted average of the two scores that tends towards the lower, thus penalizing large gaps between P and R:

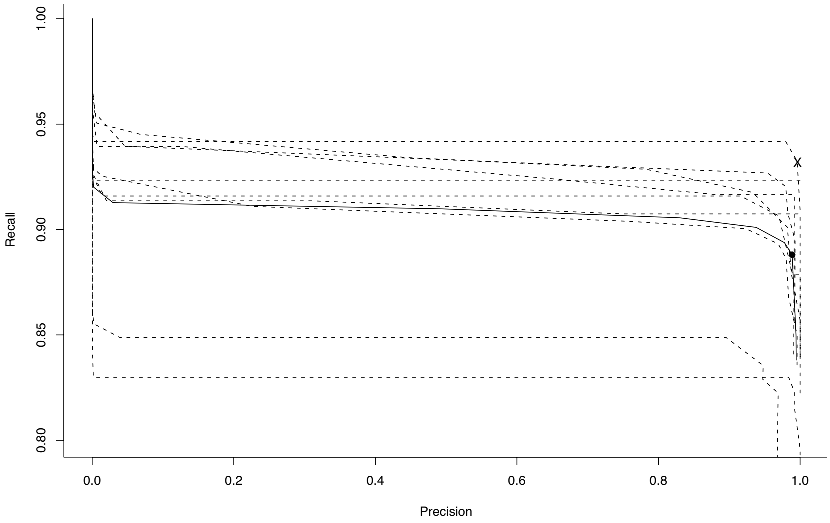


Fig. 1. Precision-recall curves (dashed lines) for the ten topics used for calibration. The single best data point (maximum F-measure, see below) is marked with an X. The solid line is the mean of the ten real curves, and the marked point indicates the maximum F-measure on this curve.

$$F = \frac{2 \times P \times R}{P + R} \tag{1}$$

These scores (up to $d = 50$) are shown in Table 5. From this table, we then picked the fuzz factor which yielded the highest average F-measure across all topics: 15, shown in bold in 5 and marked with a small circle on Fig. 1. This was then kept constant throughout all the matching runs in the following investigation.

Note that $d = 15$ represents the best fuzz factor on one of the lower-scoring calibration sets, *feverish illness*, and a close second best (after $d = 20$) on the other, *immunization.* Despite these two topics proving more difficult for our algorithm, the overall characteristics of their precision-recall curves are comparable to the others, and $d = 15$ would have been a reasonable choice of fuzz factor even if the topics in our main PsycINFO investigation had been similar to these two.

3.3 Performance

During calibration, we timed the execution of the matching script with the Linux `time` command. On a laptop with an Intel Core 2 Duo processor it took, on

Table 5. The mean precision (P) and recall (R), and hence F-measure (F), across all calibration topics, at each fuzz factor (d) from 0-50. All values are percentages. For brevity, scores for $d > 50$ are not shown; the precision and thus F-measure are practically zero in this range.

d	P	R	F
0	99.5	83.8	91.0
5	99.4	85.0	91.6
10	99.2	86.6	92.5
15	**98.9**	**88.8**	**93.6**
20	97.8	89.4	93.4
25	93.8	90.1	91.9
30	83.0	90.6	86.6
35	49.2	91.0	63.9
40	2.9	91.3	5.7
45	0.2	92.0	0.4
50	0.1	92.7	0.1

average, 107 seconds to compare two 1000-record files. This equates to almost 10,000 comparisons per second, with each string under comparison having a mean length of 80 characters (normalized length of title + first author surname). This high throughput makes it feasible to use the algorithm for larger-scale studies that would be completely impractical by hand. During the course of this experiment we found data preparation (exporting bibliographic records from Reference Manager, Access etc.) to be much more time-consuming than actually running the algorithm.

3.4 PsycINFO Analysis

Table 6 shows the results of running the tests described in the Methods section on each of the 21 real data sets from three past guideline projects, using the fuzz factor of 15 determined during calibration. For each topic, we extracted three important figures. These were: the number of hits (citations) in PsycINFO which were not present in any of the corresponding searches of MEDLINE, EMBASE, CINAHL or Cochrane; the number of PsycINFO-only hits which had been judged relevant by a reviewer, based on title, abstract etc.; and the number of resulting full-text papers which had been judged fit for inclusion in a systematic review. The results demonstrate that PsycINFO made a very small contribution to these projects, since only four of the 21 PsycINFO searches yielded any reviewable material at all, and this was limited to a single paper in three cases and three papers in the remaining case (a question about patient education). These six included papers made up only 0.3% of the grand total of 2005 papers which were included in systematic reviews for any of these guidelines. The vast majority of PsycINFO hits (96%) were weeded out as irrelevant, without the reviewers ever examining their full texts. The attrition rate at the full-text stage was slightly lower but still high, with 84% of the relevant papers being excluded for more sophisticated reasons. The reviewers' records of their reasons for excluding each paper are not comprehensive enough to allow a systematic analysis; an informal inspection suggested that there is no single predominant reason, but rather a mixture of the scenarios described in Sect. 2.4.

Table 6. For each of the 21 topics from previous guidelines, this table shows the number of: articles unique to PsycINFO, *relevant* articles unique to PsycINFO, and articles unique to PsycINFO fit for inclusion in a systematic review. The last line shows totals across all topics. For convenience, the last two columns show the relevant and included articles as percentages of the number of PsycINFO-only articles.

Guideline	Clinical question	Unique	Unique & relevant	Included in review	Percentage relevant	Percentage included
HMB	*ed*	21	0	0	0.00%	0.00%
HMB	*pe*	296	11	3	3.72%	1.01%
HMB	*peh*	35	0	0	0.00%	0.00%
HMB	*qol*	16	5	1	31.25%	6.25%
HMB	*rf*	40	2	0	5.00%	0.00%
HMB	*rfac*	2	0	0	0.00%	0.00%
IPC	*cc*	1	0	0	0.00%	0.00%
IPC	*comm*	193	7	1	3.63%	0.52%
IPC	*hvb*	3	0	0	0.00%	0.00%
IPC	*int1*	4	0	0	0.00%	0.00%
IPC	*int2*	6	3	0	50.00%	0.00%
IPC	*pain*	80	2	0	2.50%	0.00%
IPC	*pb*	2	2	0	100.00%	0.00%
IPC	*supp*	18	3	0	16.67%	0.00%
UI	*alt*	36	0	0	0.00%	0.00%
UI	*cont*	7	0	0	0.00%	0.00%
UI	*life*	185	2	1	1.08%	0.54%
UI	*pbt*	25	1	0	4.00%	0.00%
UI	*pharm*	6	0	0	0.00%	0.00%
UI	*ques*	3	0	0	0.00%	0.00%
UI	*tens*	1	0	0	0.00%	0.00%
All	**All**	**980**	**38**	**6**	**3.88%**	**0.61%**

To put these results in perspective, we compared the total number of hits across all databases with the numbers found relevant and ultimately included in a review. Averaged across all the 21 data sets, we found that the rate of attrition was less than for the PsycINFO results at both the weeding and exclusion stages. Overall, 91% of database hits were discarded as irrelevant, but only 71% of those remaining failed to meet the inclusion criteria for systematic reviewing based on their full text. This demonstrates that PsycINFO searches are noisier than average (greater proportion of irrelevant hits) and return less useful studies (greater proportion of excluded papers), at least on the subjects covered by the NCC-WCH.

4 Conclusions

The results of this investigation have allowed us to quantify, albeit crudely, the return on the NCC-WCH's investment into PsycINFO. Given that the 21 search result sets examined in the previous section represent the principal fruits of a year's subscription (£1634 in 2005 when these searches were carried out), and yielded only six reviewable papers, this represents a financial cost of over £272 just to identify each paper. A fourth guideline from the same year also searched PsycINFO for three of its clinical questions, returning a total of 20 raw hits, but unfortunately detailed search logs no longer exist for that project. Extrapolating from the results of our experiment, however, suggests it is likely that none of these papers were eventually reviewed, since in the three guidelines tested PsycINFO yielded 1275 raw hits but only six that were ultimately reviewed (less than one per 200 hits).

The cost per paper arising from the PsycINFO subscription fee is almost certainly dwarfed by the less easily quantifiable costs of building PsycINFO searches, downloading and managing the results, manually weeding out irrelevant hits, ordering and obtaining full papers, and critically appraising each one for eligibility. The value of these activities is called into question when less than one percent of citations found uniquely in PsycINFO will be used as evidence. Idealistically, one might argue that a truly systematic review should include all available evidence whatever the cost, but resources are finite and might be more effectively spent elsewhere. For comparison, EMBASE cost the NCC-WCH £4308 during 2005, and CINAHL £1395. The Cochrane library and MEDLINE are free via their standard web interfaces, and an Ovid-based subscription to these is available to the NCC-WCH for free via its parent organization (the Royal College of Obstetricians and Gynaecologists). Empirical experience has taught us that the vast majority of database hits come from MEDLINE and/or EMBASE, but the time and resources were not available to quantify this precisely.

We have shown, then, that for the areas of medicine which the NCC-WCH specializes in, PsycINFO offers sparse coverage, noisy results and content of limited evidential value. Of course, one must not extrapolate from our results to conclude that PsycINFO is necessarily unsuitable for use in evidence-based research. All of the 21 questions in our test set related to physical conditions or interventions with a psychological aspect, but a review or guideline in the field of mental health would almost certainly find more usable material in PsycINFO. Furthermore, the exact criteria for excluding a study vary from review to review, both within a research group and between groups, and are influenced by various extrinsic factors such as resource issues and availability of other evidence. For example, we generally weed out articles in foreign languages, due to the time and cost implications of inter-lingual reviewing, unlike the Cochrane Collaboration who encourage their reviewers to have relevant papers translated whenever possible [16]. Rather than implying any general message about PsycINFO, we hope we have demonstrated that data-driven studies of medical literature resources are tractable with the aid of text mining techniques, and have a role in process improvement and strategic decision-making.

4.1 Future work

Two further questions one could ask of the results of this investigation are: how would the results of the systematic reviews based on the four searches which found PsycINFO-only hits have been affected by the removal of these papers, and what consequences for cost-effectiveness analysis would this have? In other words, how clinically and economically significant were these six papers? These are questions that can only be answered via additional analyses by systematic reviewers and health economists respectively, and are thus beyond the scope of an informatics paper. However, it should be noted that all of the NCC-WCH's guidelines are developed in collaboration with a committee of subject experts, and are put out for consultation to an open panel of stakeholder organizations and individuals before publication. Both of these groups have the opportunity to alert the reviewers to papers and other data that have been missed by database searches, thus providing a safety net for important pieces of evidence that might change the results of a review.

From the technical point of view, there are several additional avenues along which this work could be extended. There are many other ways to compare two strings non-exactly. One class of solutions comes from bioinformatics, where local alignment (identifying regions of similarity or identity) and global alignment (finding the best match that takes in the entire spans of the strings) are indispensable methods for the analysis of gene, protein and RNA sequences. In addition, there are various techniques developed for the comparison of unsegmented citations that could be evaluated for their performance on the already-segmented records in our databases, some of which are discussed in Sect. 1.4. Since this project was launched not as a research effort but as a practical investigation into the efficiency of the NCC-WCH's business process, we did not have the time or resources to benchmark a number of different solutions to the same problem. Neither were we able to perform a detailed Precision/Recall error analysis in the time available. However, since we have demonstrated an unsupervised method for obtaining suitable data sets for this class of problem (Sect. 2.2), we hope that we might inspire interested investigators to perform further experiments in this important area.

References

1. Chaudhari, T., McGuire, W.: Allopurinol for preventing mortality and morbidity in newborn infants with suspected hypoxic-ischaemic encephalopathy. Cochrane Database of Systematic Reviews **Issue 2** (2008)
2. Wilkins, T., Gillies, R.A., Davies, K.: EMBASE versus MEDLINE for family medicine searches: Can MEDLINE searches find the forest or a tree? Canadian Family Physician (2005) 848–849
3. Suarez-Almazor, M.E., Belseck, E., Homik, J., Dorgan, M., Ramos-Remus, C.: Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Controlled Clinical Trials **21** (2000) 476–487

4. Sampson, M., Barrowman, N.J., Moher, D., Klassen, T.P., Pham, B., Platt, R., John, P.D.S., Viola, R., Raina, P.: Should meta-analysts search EMBASE in addition to MEDLINE? Journal of Clinical Epidemiology **56** (2003) 943–955
5. Royle, P., Milne, R.: Literature searching for randomized controlled trials used in cochrane reviews: Rapid versus exhaustive searches. International Journal of Technology Assessment in Health Care **19** (2003) 591–603
6. McDonald, S., Taylor, L., Adams, C.: Searching the right database: A comparison of four databases for psychiatry journals. Health Libraries Review **16** (1999) 151–156
7. Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity uncertainty and citation matching. In: Advances in Neural Information Processing Systems 15, MIT Press (2002) 1425–1432
8. Wellner, B., McCallum, A., Peng, F., Hay, M.: An integrated, conditional model of information extraction and coreference with application to citation matching. In: AUAI '04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press (2004) 593–601
9. Lawrence, S., Giles, C.L., Bollacker, K.D.: Autonomous citation matching. In: AGENTS '99: Proceedings of the Third Annual Conference on Autonomous Agents, ACM (1999) 392–393
10. Monge, A.E., Elkan, C.P.: An efficient domain-independent algorithm for detecting approximately duplicate database records. In: Proceedings of the SIGMOD 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery, ACM (1997) 23–29
11. Goode, P.S., Burgio, K.L., Locher, J.L., Roth, D.L., Umlauf, M.G., Richter, H.E., Varner, R.E., Lloyd, L.K.: Effect of behavioral training with or without pelvic floor electrical stimulation on stress incontinence in women: A randomized controlled trial. Journal of the American Medical Association **290** (2003) 345–352
12. Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press (1997)
13. National Collaborating Centre for Women's and Children's Health: Diagnosis and management of heavy menstrual bleeding. National Institute of Clinical Excellence, UK (2007)
14. National Collaborating Centre for Women's and Children's Health: Intrapartum care—management and delivery of care to women in labour. National Institute of Clinical Excellence, UK (2007)
15. National Collaborating Centre for Women's and Children's Health: Urinary incontinence—the management of urinary incontinence in women. National Institute of Clinical Excellence, UK (2006)
16. Clarke, M.: personal communication. Director, UK Cochrane Centre (2008)

Data Mining of Clinical Oral Health Documents for Analysis of the Longevity of Different Restorative Materials in Finland

Taina Käkilehto[1], Sinikka Salo[1] and Markku Larmas[1,2]

Department of Pedodontics, Cariology and Endodontics, Institute of Dentistry,
University of Oulu[1], and Oulu University Hospital[2]
Taina Käkilehto
Institute of Dentistry
University of Oulu
P.O.Box 5281
FIN-90014 University of Oulu, Finland
Tel.+358-8-5375011; Fax +358-8-5375560
E-mail taina.kakilehto@pp.inet.fi

Abstract. Evidence based dentistry has shown that different restorative materials have different survival times. Our null-hypothesis is that data mining technique and a practice-based dentistry approach analysed in a scientifically sound way from normal dental records should reveal this. Dental records from 1906 patients and altogether 19,892 restorations in three Finnish age cohorts were analysed. Survival curves (Kaplan-Meier) for each of the restorative materials were drawn. Median survival times for amalgam and resin based composites were over fifteen years in older cohorts. More than 60% of the silicate cement restorations were replaced within five years, and more than 50% of the glass ionomers within seven years. There was a significant reduction in the longevity of amalgams in the cohort 1980. Data mining of digital oral heath documents would be a useful tool to analyze survival curves of new restorative materials on practice-based manner in real life conditions.

1 Introduction

The longevity of dental restorations is an important health concern for the patient, the dentist, and the various forms of insurance systems. Only longitudinal studies are appropriate to give an exact insight in the longevity of restorations.[1]. The evolution of clinical evidence of efficacy normally evolves from laboratory studies through animal studies to case series, to controlled clinical trials, and finally systematic reviews and forms the evidence-based dentistry concept [2]. The knowledge of the longevity of dental restorations is retrospective, and is normally based on surveys of materials used for different clinical trials, which have been carried out under optimal conditions [3]. However, Good Clinical Practice: that is standardized placement of restorations by calibrated operators into standardized patients for a follow-up of decades, is not possible in practice for dental restorations [2]. It is logical to proceed from the science–based approach to research protocols implemented to practice based research. This would link science to real life conditions in developing dental materials [4]. The results of longevity of restorations can be presented in different ways, but in the survival analysis it is necessary that the censored data are taken into account. By using the Kaplan & Meier method (1958) all observations, including the censored ones, are used [1].

When studying the lifetimes of dental restorations, one is, faced with the fact that each patient may contribute multiple survival data even for the same restorative materials [5]. This creates a dependence problem because standard survival analysis requires independent data, which is not possible inside the same oral cavity [5]. Another issue is the fact that each dentist will also contribute a non-standardized operation method and therefore, comparisons between materials in Good Clinical Practice should be dentist-specific or conducted by standardized operators, which is very expensive.

The aim of this study was to test on a large scale if data mining based on normal dental records can be used for scientific analysis of the longevity of dental restorations in practice. Normal paper-based oral health documents from public health centres before the "digital era": that is before the implementation of electronic patient record, were collected, after which a digital intermediate file was formed for computer analysing of longevity of dental restorations. The Kaplan-Meier curves could be created individually for different dental materials on different tooth surfaces for comparisons of results.

The primary working hypothesis is that data mining can be used for scientific analysis of normal oral health care records and different restorative materials have different survival times on different teeth groups. The aim was also to study if there were any difference between maxillary vs. mandibular teeth or left and right sides of the patients' jaws.

2 Materials and Methods

Public dental records were found to be reliable when compared with those made by trained research team [6]. Because electronic patient records were taken into practice at a certain time at each of the health centres in Finland, we modified the paper documents into an electronic from patient records before `digital era` for analysis. Originally paper-based documents was individually aggregated into a digital intermediate file in a network of four public dental health centres in four towns in the north of Finland: Kemi (24,000 inhabitants), Oulu (110,000), Raahe (18,500) and Tornio (24,000). The towns had similar demographic structure and the same basic dental restorative treatment systems, with identical dentist to population ratios and examination intervals during the observation time 1965- 1995.

The Committee of Ethical Affairs of Oulu University Hospital and each health centre were involved in granting permission for this study at 9.4.2001. The data were collected from copies of the oral health documents. In the copies the last names and social security codes were obliterated in order to make the data anonymous for data handling and analysing processes.

Patient records were collected from three cohorts: subjects born in 1960- 63 (1960 cohort), subjects born in 1970-71 (1970 cohort) and subjects born in 1980-81 (1980 cohort). The subjects were randomly selected from the files of the health center of Oulu in age cohorts 1970 and 1980. In Kemi, Tornio and Raahe the subjects represent the whole age-cohort (Table1).

Table 1 Number of subjects with dental restorations in different health centers and study cohorts

	Cohort	Kemi	Oulu	Tornio	Raahe	All
The whole study material	Born 1960	194	-	93	-	287
	Born 1970	120	281	117	193	711
	Born 1980	186	369	178	175	908
Subjects with one or more restorations	Born 1960	193	-	93	-	286
	Born 1970	116	273	116	184	689
	Born 1980	126	255	145	125	651

Total number of patient records was 1906. The primary criterion for inclusion in the analysis was that the subjects had one or more restorations. In the 1960 cohort 1 patient (0,3%), in the 1970 cohort 22 patients (3,1%) and in the 1980 cohort 257 patients (28,3%) had sound dentition. Altogether 1626 subjects were included (Fig 1). A digital intermediate file containing the following information was compiled from the patient documents: date of birth, gender, cavitated carious lesions and/or restorations on each tooth surface and extracted or missing teeth. Observations were made on every tooth except third molars.

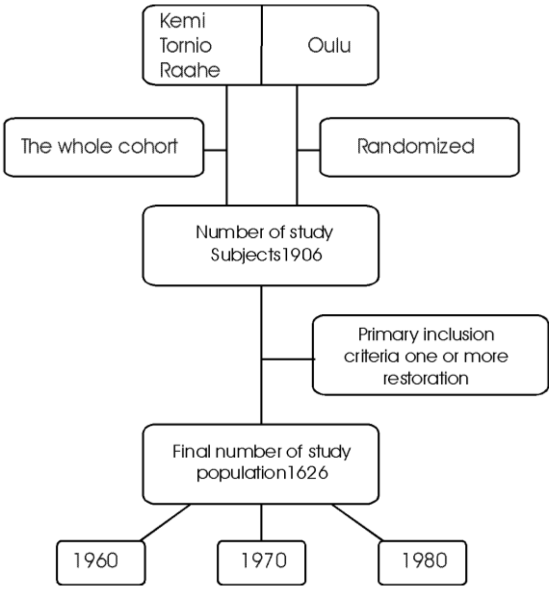

Kemi
Tornio
Raahe
Oulu
The whole cohort
Randomized
Number of study Subjects1906
Primary inclusion criteria one or more restoration
Final number of study population1626
1960
1970
1980

Fig. 1 The study design

The children received their annual dental examinations and treatment when needed in health centres free of charge from birth until 19 years of age. All patients were born and living in the town in question. The dental records covered periods ranging from 5 years of age up to past the 25th birthday, the information being recorded at annual examinations in the health centres. The dates of the placement of dental restorations (recorded by codes for amalgam, composites, glass ionomers, and silicate cements) of each individual tooth/tooth surface from the progress notes of dental records were determined. Altogether 19,892 restorations were placed during the observation period 1965-1995. The replacement of restorations was deemed to have occured when an old restoration was replaced on the same restored surface. It was the beginning the follow-up for that restoration. The time from the first placement to the replacement of the restoration represented the "survival time" or longevity of each dental restoration.

2.1 Treatment of data and statistical analysis

The ages of the subjects and dates of operations were entered into the computer to an accuracy of 1 day, and the date of the annual examination to an accuracy of 1 month. Dental chart markings and progress notes were coded for the computer [7]. The methods of survival analysis were employed, using the Kaplan-Meier method for each tooth/surface and restorative material being evaluated by the product-limit method

When the placement time for restorations is not simultaneous, the data are called left censored. On the left side of the time scale, the date of restoration placement was recorded for the origin of the follow-up of particular restoration. On the right scale the “fail” (replacement) of the restoration was recorded as the end of survival, but at the same time as a new origin for the new follow-up of a new restoration.

In follow-up studies the restorations may not fail and the “real” survival time is not found, but a method of estimating the x year survival change in which all information is used is the procedure of Kaplan Meier. So, on the right side of the time scale, the follow up of restorations was right-censored at the last examination or treatment visit. [1], [8].

In the analysis, the SAS statistical software program, the CIA program of Gardner and Altman and the SPSS program were used for drawing the survival curves. The Median survival time (MST) was calculated for each survival curve [8].

The statistical differences of the survival curves were compared using the log-rank test. A difference was considered to be statistically significant if the p-value was <0,05.

3 Results

No differences were observed in the longevity of restorations when all teeth were combined between the left and right side or between maxilla and mandible (Fig 2.), also found earlier [9].

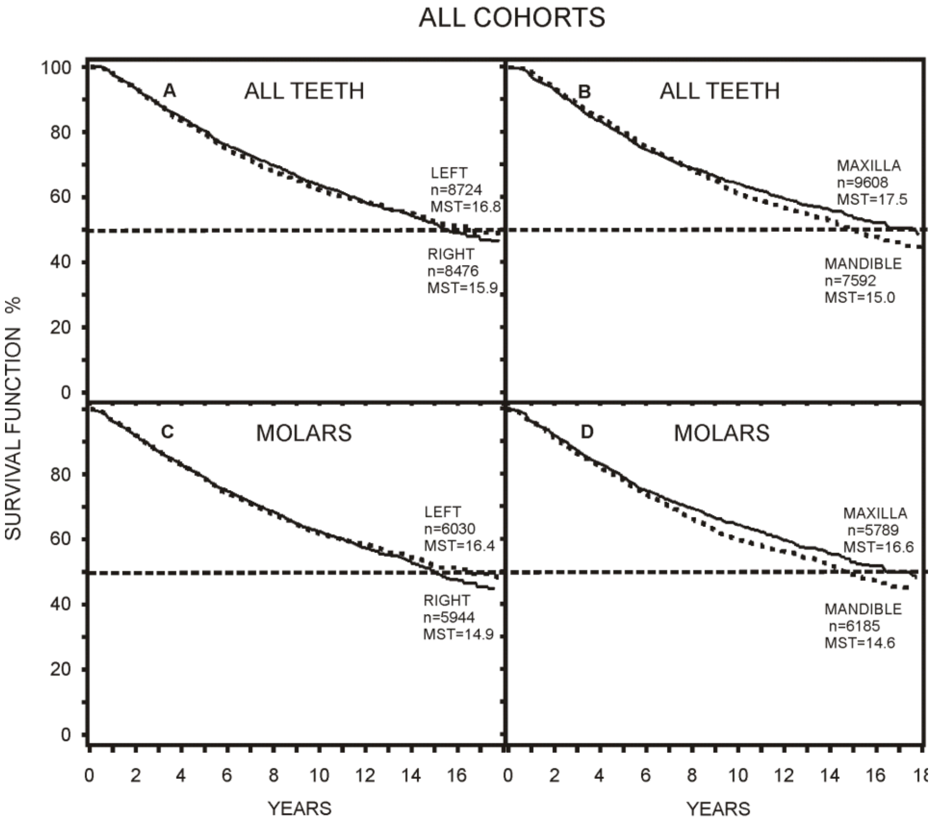


Fig. 2 Kaplan-Meier survival curves for restorations with mean survival times for all teeth (A,B) and molar teeth (C,D) in left and right sides of the jaws and in maxilla and mandible.

The 1980 cohort had shortest survival times with regard to all materials analyzed The difference between survival times were statistically significant (p< 0,05) (Fig 3). Silicate cement restorations were placed in the 1960 cohort only. When restoration on occlusal surfaces were analyzed separately, no big differences were found between the 1960 and 1970 cohorts with amalgam restorations, however, no composite restorations were placed on occlusal surfaces in the 1960 cohort. The MST for amalgams on occlusal surfaces was 16.8 years in the 1960 cohort, 13.6 years in the 1970 cohort, but only 7.9 years in the 1980 cohort. Glass ionomers and composites had MST from 4.9 to 7.3 years on occlusal surface in the 1970 and 1980 cohorts (Fig. 3).

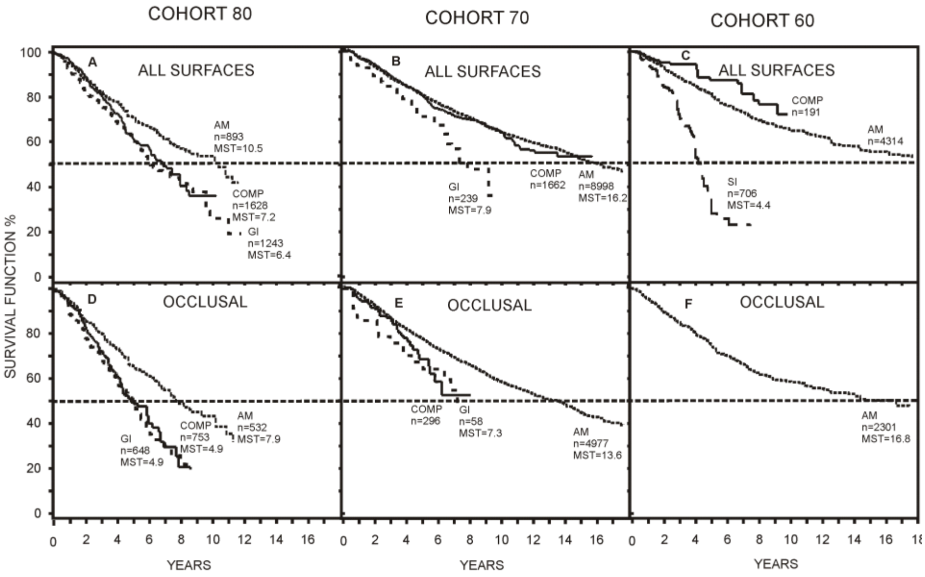


Fig. 3 kaplan-Meier survival curves for cohorts 1960, 1970 and 1980, on all tooth surfaces (Upper panel) and occlusal surfaces (lower panel) with MST (when could be counted) for amalgam (AM), composites (COMP), glass ionomers (GI) and silicate cements (SI)

When all cohorts were combined, amalgam was observed to be superior to any other restorative material in teeth analysed (Fig. 4). Composites have shorter longevity (Fig. 4B) than amalgam but longer than glass ionomer (Fig. 4 C) in incisors, premolars and molars, which always had the shortest survival (Fig. 4). More than 60% of the silicate cement restorations were replaced within five years, and more than 50% of the glass ionomers within seven years (Fig. 4).

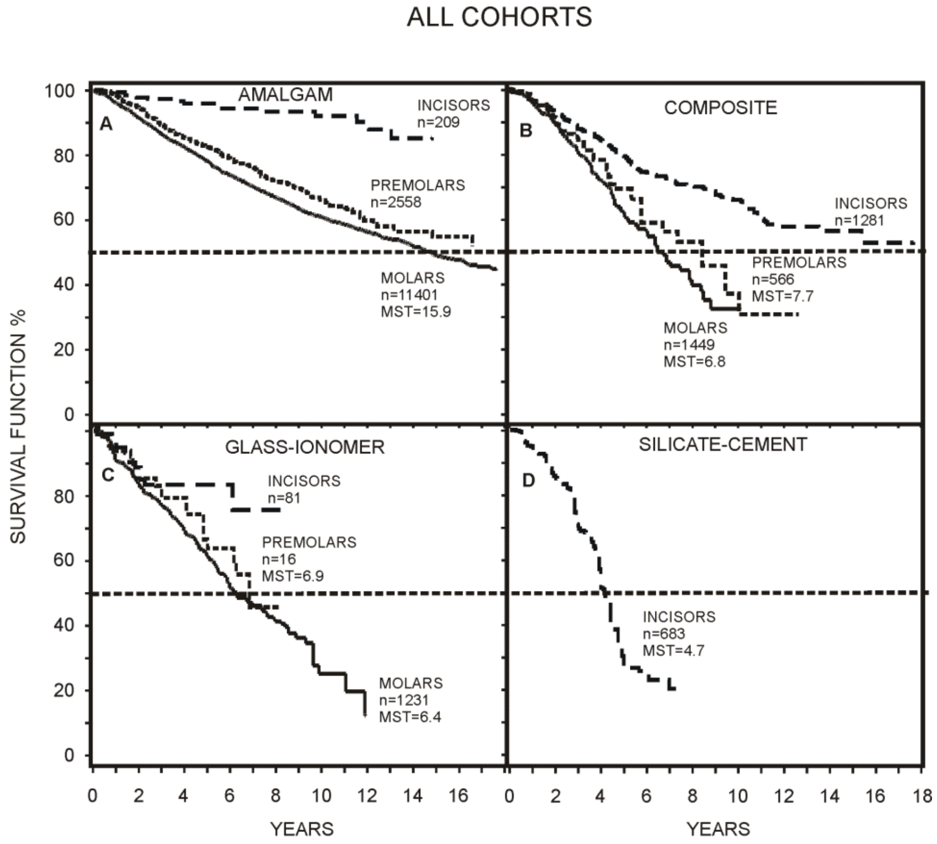


Fig.4 Kaplan-Meier survival curves for amalgams, composites, glass ionomers and silicate cements for incisors, premolars and molars with MST values

When the 1970 and 1980 cohorts were compared, it was observed that amalgam restorations in the 1980 cohort had shorter longevity (MST 10.7 years) than in the 1970-cohort (MST 15.8 years) in molars (Fig. 5A, D) In the 1970 cohort, some amalgam restorations (altogether 101) were placed on incisors, but premolars and molars harboured most of the amalgam restorations (Fig. 5A, D). No amalgam restorations were made in premolars or incisors in the 1980 cohort. The longevity of composites was shorter in the 1980 cohort than in the 1970 cohort in all teeth analysed (Fig. 5B, E), and the survival was shortest with glass ionomers in all these groups (Fig. 5C, F).

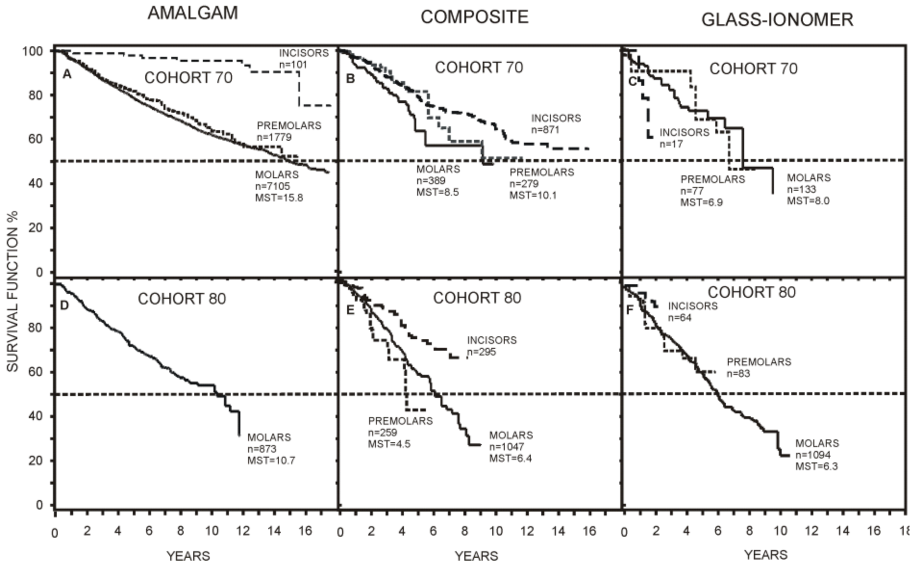


Fig.5 Kaplan-Meier survival curves for cohorts 1970 (upper panel) and 1980 (lower panel) for amalgams, composites and glass ionomers. Note that the number of amalgam fillings has reduced to 873 in molars and to zero in premolars and incisors in the 1980 cohort, while they were 7105 on molars in the 1970 cohort

4 Discussion

Data mining is a discipline using modern computing tools to solve problems and it´s methods may provide advantages over traditional statistical methods in dental data. [10]. In this study, we compared the longevity of restorative materials between three decades from 1965 to 1995. Because electronic patient records were taken into practice at a certain time at each of the health centres in Finland, we modified the paper documents into an electronic from patient records before `digital era` for analysis. The digital intermediate file was compiled by two calibrated dentists, and thus, the quality and reliability of data is considered to be high for scientific purposes. Already earlier, database compiling was observed to be a convenient way of analyzing the normal electronic patient record files from the health centers and for scientific purposes in a practice-based research fashion to compare observations made in real-life conditions on findings of epidemiological studies [11].

In this study the survival times of amalgam and composite dental restorative materials were about the same when the data set was analysed as a whole, but more detailed analysis revealed differences: amalgam restorations survived longest especially on occlusal surfaces where the survival time of the other materials was shorter. In addition, they were much shorter with glass ionomers (Fig. 1), as also seen in the British analyses [12]. The short longevity of silicate cement restorations in the 1960 cohort resulted in their withdrawal from the markets already in the 1970`s in

Finland as well as in all other industrialised countries. The sharp drop of the survival curve of silicate cements showed us immediately that the material was urgently replaced. The need for an effective clinical reporting system was obvious at that time [2].

The longevity of composites was also shorter in the 1980 cohort than in the 1970 and 1960. This may be due to rapid, progressive development of new generations of resin-based composites without documented efficacy prior to their use in practice. These restorations resulted in early failures caused by a variety factors for example the unexpected technique sensitivity associated with their placement [2].

Mjör et al. divided the factors which affect the longevity of dental restorations into three categories: 1) the clinician, 2) the restorative material and 3) the patient [2]. During the past 50 years, the changes in the practice of restorative dentistry have focused on the development of new restorative materials and clinical techniques, usually without any documentation [2]. In this study, the longevity of amalgam restorations was shorter in the 1980 cohort than in the 1970 and 1960, where they were about identical (Fig. 3). This may indicate that the public demand for tooth coloured restorations and public discussion of the possible harmful effects of amalgam mercury led patients to demand the change of amalgam restorations to tooth coloured restorations, as also discussed in the British survey [12]. The increased occurrence of tooth coloured restoration of the 1980-cohort also supports this view.

When were compare our median survival values to those recently published in Finland, our survival times for glass ionomers, composites and amalgams were 25 % to 50 % longer than the previously published result [13]. The difference may be explained by the methodological differences. In order to evaluate the reasons of failed restorations, Forss and Widström asked dentists to record information for each restoration they placed during a three-day period. The authors compared the survival of failed restorative materials using the median values. In our study, a retrospective, practice based study was conducted after measuring all placed restorations from dental records, and the Median survival time (MST) was calculated for each survival curve.

Comparisons to other recent practice based studies revealed that the variation between survivals of different materials is large. In the the U.S.A , more than 90 % of both composites and amalgams survived at 5 years with the patients of Washington Delta Service if the initial and follow-up dentist was the same. This fell to 70 % when the patient went to a different dentist [14]. The Life span of restorations by treatment type in in adults in Great Britain varied between 62-72% for amalgams, 58% for resin composites and 53% for glass ionomers at 5 years [12], [15]. Survival of amalgam restorations among Royal Air Force personnel in Great Britain was also about the same at 5 years, and the single level 50% survival was exactly 11.5 years and multilevel 50% 12.5 years [16], [17]. The corresponding figures in Finland are slightly better (50% being over 15 years) than those in Britain, but in both counties they are worse than those in the United States if the treating and follow up dentist remained the same. If the US figures are compared to the European data, indicating that dentist related factors are probably more important than the actual restoration material in the longevity of restorations. This will be analyzed later in detail with this dataset.

In the 21st century restorative materials are still under great development without effective clinical reporting systems. Data mining of digital oral heath documents would be a useful tool to analyze survival curves of new restorative materials in a practice-based manner in real life conditions. It would give valuable feedback to clinical practice more easily compared to the research initiated and performed in a laboratory.

5 Conclusion

This study clearly indicates that data mining of digital oral heath documents would be a useful tool to analyze survival curves of new restorative materials in a practice-based manner in real life conditions. The survival time of restorative materials was different in different teeth groups but no differences were found when comparisons were made between the mandible and the maxilla or left and right sides of the jaws. Median survival times for amalgam and resin based composites were over fifteen years in older cohorts in Finland. There was no change in the longevity of amalgams between the cohorts 1960 and 1970 but a significant reduction in the 1980 cohort suggesting that patient related factors in the selection of restoration materials was more important than before.

References

1. Leempoel P.J.B., Van`t Hof M.A., De Haan A.F.J.: Survival studies of dental restorations: criteria, methods and analyses. J Oral Rehabil. 16, 387--394 (1989)
2. Mjör I.A., Gordan V.V., Abu-Hanna A., Gilbert G.H.: Research in general dental practice. Acta Odontol Scand. 63,1--9 (2005)
3. Jokstad A., Bayne S., Blunck U., Tyas M., Wilson N.: Quality of dental restorations. Int Dent J. 51, 117--158 (2001)
4. Niederman R., Leitch J.W.: "Know what" and "know how": knowledge creation in clinical practice. J Dent Res. 85, 296--297 (2006)
5. Aalen O.O., Bjertness E., Sonju T.: Analysis of dependent survival data applied to lifetimes of amalgam fillings. Stat Med.14, 1819--1829 (1995)
6. Seppä L., Hausen H., Pöllänen L., Helasharju K., Kärkkäinen S.: Past caries recordings made in public dental clinics as predictors of caries prevalence in early adolescence. Oral Epidemiol. 17:277—281 (1989)
7. Larmas M.A., Virtanen J.I., Bloigu R.S.: Timing of first restorations in permament teeth: a new system for oral health determination. J Dent. 23, 347--352 (1995)
8. Dawson-Saunders B., Trapp R.G.: Basic and Clinical biostatistics 2. ed, 186--206 (1994)
9. Drake C.W.: A comparison of restoration longevity in maxillary and mandibular teeth. JADA. 116, 651--654 (1988)
10. Gansky S.A.: Dental Data Mining: potential Pitfalls and Practical Issues. Adv Dent Res 17, 109 --114 (2003)
11. Korhonen M., Salo, S., Suni J., Larmas M.: Computed online determination of life-long index values for carious, extracted, and/or filled permanent teeth. Acta Odont Scand. 65, 214--218 (2007)

12. Lucarotti P.S.K., Holder R.L., Burke F.J.T.: Outcome of direct restorations placed within the general dental services in England and Wales (Part 1): variation by type of restoration and re-intervention. J Dent. 33, 805--815 (2005)
13. Forss H., Widström E.: From amalgam to composite: selection of restorative materials and restoration longevity in Finland. Acta Odontol Scand. 59, 57--62 (2001)
14. Bogacki R.E., Hunt R.J., delAquila M., Smith W.R.: Survival analysis of posterior restorations using an insurance claims database. Oper Dent. 27, 488--492 (2002)
15. Lucarotti P.S.K., Holder R.L., Burke F.J.T.: Analysis of an administrative database of half a million restorations over 11 years. J Dent 33, 791--803 (2005)
16. Giltrope M.S., Mayhew M.T., Bulman J.S.: Multilevel survival analysis of amalgam restorations among RAF personnel. Community Dental Health. 19, 3--11 (2002)
17. Manda S.O.M., Gilthrope M.S., Tu Y.-K., Blance A., Mayhew, M.T.: A Bayesian analysis of amalgam restorations in the Royal Air Force using the counting process approach with nested frailty effects. Stat. Meth Med Res. 14, 567--578 (2005)

Towards Resource-Efficient Construction of a Full Parser for Finnish Nursing Narratives

Veronika Laippala[1,3], Filip Ginter[1], Sampo Pyysalo[2], and Tapio Salakoski[1,2]

[1] Department of Information Technology
[2] Turku Centre for Computer Science (TUCS)
[3] Department of French Studies
20014 University of Turku, Finland
first.last@utu.fi

Abstract. In this paper, we present a formal grammar and a parser for the language used in daily nursing notes in a Finnish Intensive Care Unit (ICU). We analyze ICU Finnish as a sublanguage, identifying its specific features that facilitate for example the development of a specialized grammar. The identified features include frequent omission of finite verbs, limitations in allowed syntactic structures, and domain-specific vocabulary.

The grammar is implemented in the LKB system in a typed feature structure formalism. The lexicon is automatically generated based on the output of the FinTWOL morphological analyzer adapted to the clinical domain; the grammar thus efficiently uses existing resources for Finnish. The grammar currently covers 67% of ICU Finnish sentences, producing highly accurate best-parse analyzes with F-score reaching 93%. We find that building a parser for the highly specialized domain sublanguage is not only feasible, but can also be done in a remarkably resource-efficient manner, given an existing morphological analyzer with broad vocabulary coverage. The resulting parser enables deeper analysis of the text which was previously not possible.

1 Introduction

The potential of natural language processing methods applied to clinical text has long been recognized, with a number of important applications in decision support and patient management, mining of trends and correlations, patient profiling, etc. While a purely statistical treatment suffices for tasks such as document classification, methods aiming towards deeper analysis of the text in the form of limited text understanding rely on some form of syntactic analysis as a critical processing step.

Syntactic analysis provides an account of sentence structure, revealing the syntactic roles of individual sentence constituents as well as their mutual relationships. Depending on application, different parsing strategies can be employed, providing different levels of detail. The METAMAP system [1], for example, incorporates a partial parser that recognizes noun and verb phrase boundaries, but does not analyze their mutual relations.

In contrast to the limited analysis provided by partial parsers such as the one used by METAMAP, full syntactic parsers recover not only the constituents of the sentence but also fully resolve their mutual relationships, thus giving a considerably more detailed account of sentence structure (see Figure 1 for an example). Methods based on full parsing have recently been gaining in popularity, largely due to the substantial progress in statistical parsing. Modern statistical parsers trained on large treebanks (see e.g. [2]) are sufficiently robust, accurate and computationally efficient while providing a full parse.

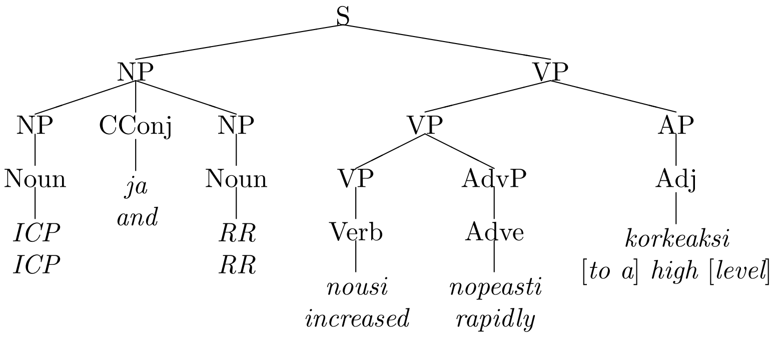


Fig. 1. Example parse tree for the Finnish sentence *RR ja ICP nousi nopeasti korkeaksi* (Eng. *RR and ICP increased rapidly to a high level*).

In this paper, we describe the development of a full parser for daily nursing notes from an intensive care unit (ICU) in a Finnish hospital to support the application of natural language processing methods in the domain of nursing narratives. The ICU narratives are unstructured documents written by nurses as a daily record of the condition of each patient during their work shift and serve as a crucial source of information about the status of the patient. Many possible applications of language processing methods in this domain (see, e.g., [3]) would benefit from the availability of a full syntactic analysis of the text. The proposed parser can thus be seen as a core resource on top of which further applications can be developed. As an example, an application of interest would be the ability to summarize the notes by extracting trends w.r.t. important topics such as hemodynamics, oxygenation, etc.

There are several ways in which a full parser can be developed. Whenever a sufficiently large treebank is available, it is possible to train a statistical parser, as has been demonstrated, for example, by Sagae et al. [4] for biomedical English. Unfortunately, as is often the case for minority languages lying outside of the main-stream interest of the NLP community, we do not have a sufficiently large treebank at our disposal, neither for general Finnish, nor for the sublanguage of intensive care nursing narratives, referred to as *ICU Finnish* throughout this paper. Therefore, the approach of training a statistical parser is not applicable without a significant effort invested into first developing such a treebank.

As an alternative to training a statistical parser, it would be possible to adapt a general language, hand-written, rule-based parser to the specific sublanguage of interest. The only available broad-coverage full parser of Finnish, the Connexor Machinese Syntax[4], is, however, a complex closed-source parser whose adaptation is infeasible outside of the commercial entity.

Since the other methods are not applicable, building a rule-based parser specifically for ICU Finnish is basically the only possibility to obtain a syntactic analyzer for it. Very broadly, a rule-based parser, or the formal grammar on which it is based, consists of two components: a lexicon and a set of syntactic rules. The *lexicon* encodes the knowledge of the words of the language and their properties, while the grammar rules encode the allowed syntactic structures. In this paper, we propose an approach in which an existing general Finnish lexicon is extended with domain vocabulary, a procedure that is not particularly resource-intensive, while the set of grammar rules is developed anew, taking advantage of the highly restricted syntax of the ICU Finnish sublanguage. This approach allows us to re-use the lexicon which ICU Finnish has in common with general Finnish and to develop a full parser for ICU Finnish with relatively little effort, thus demonstrating the viability of our approach in situations where a general language lexicon is available but not the set of formal grammar rules.

In our case, the lexicon we build our parser on is provided as part of the accurate Finnish morphological analyzer FinTWOL[5] [5] by Lingsoft Inc. FinTWOL is a de-facto standard account of Finnish morphology, analyzing each wordform into its lemma and a set of tags from a structured tagset with a number of categories forming each tag, expressing morphological information such as the primary part-of-speech (POS), tense, number, person, possessivity, etc. An example FinTWOL analysis is shown in Figure 8. Since Finnish is a highly inflective language with complex morphology, the ability to re-use an existing morphological analyzer is a crucial condition for the feasibility of our study; developing a new Finnish morphological analyzer would be extremely resource-intensive.

In order to develop a parser for a particular sublanguage, it is necessary to analyze this sublanguage and identify its specific features affecting the grammar development. Clinical, medical, and biomedical sublanguages of English have earlier been extensively studied by Friedman et al. [6, 7], among others. For Finnish home care narratives, an analysis has been given by Karvinen [8]. The results, however, are not applicable for our study as the approach and goal of the study are different and the sublanguage of home care narratives does not correspond to ICU Finnish. For Finnish ICU nursing narratives specifically, no applicable analysis is available and therefore we perform one as part of this study.

The paper is structured as follows. First we analyze ICU Finnish as a sublanguage of general Finnish, exposing the systematic differences that necessitate the development of a specialized parser. Then we introduce important aspects of

[4] http://www.connexor.eu

[5] http://www.lingsoft.fi

the technical implementation of the grammar such as the unification framework in which the grammar is developed and the treatment of out-of-lexicon wordforms. Finally, we present a preliminary evaluation of the parsing coverage and accuracy on a manually annotated treebank of ICU Finnish.

2 Theory of Language and the sublanguage analysis

ICU Finnish differs from standard Finnish significantly (Fig. 2); it is telegraphic with fragmentary sentences and frequent misspellings and abbreviations. Typographic symbols are used to replace actual words and the vocabulary contains clinical terms. As these specific features affect the automatic analysis of the language and the development of a formal grammar for it, a necessary first step is to carefully analyze these properties. A widely adopted [6, 9–12] theoretical framework for the linguistic analysis of a language for NLP applications is the Theory of Language and Information and the sublanguage theory of Harris [13, 14]. This framework enables the study of language on the semantic, lexical and syntactic levels and allows a formal and systematic analysis whose results are suitable for the development of automatic text processing systems. Moreover, the sublanguage theory of Harris also provides a framework for a detailed description of the lexico-syntactic patterns possible in the sublanguage under study.

yövuoro
Aloitetaan heng.harj.Bennetillä. Hapetus ok. saturatio 90. Putkesta ei nyt tullut nestettä.
hemodyn: diuresi toimii paine tosin välillä >100. Pulssi on rauhallinen.
tajunta hereillä, ottaa katsekontaktin. Vas. kättä jaksaa liikuttaa.
O M A I S E T: veli soittanut&käynyt. Puhunut fysioterapeutin kanssa.
2006-00-00 00:00
pitkä aamuv
HENGITYS: heikkoa, CO2 edelleen korkea. Ahdistunut mutta rauhoittuu
itsekseen. Co-operoi.
Klo 20 jälk hb 85, annettu 3 ps —> 121. hR silti 120.
Direesi niukkaa.
Muuta: Illan LÄÄKE ei annettu, tarvittaessa voi antaa.

nightshift
Starting breath.exerc.with Bennet. Oxygen ok. saturation 90. No liquid has come from the drain now.
hemodyn: diuressis is working pressure however from time to time >100. Pulse is calm.
consciousness awake, takes eye contact. Can move l. hand.
R E L A T I V E S: brother called&visited. Talked with the physiotherapist.
2006-00-00 00:00
long mornings
BREATHING: weak, CO2 still high. Anxious but calms down by himself. Kooperates.
After 20 o'clock hb 85, 3 u given –> 121. hR still 120.
Diresis narrow.
Other: Evening DRUGNAME not given, it can be given if needed.

Fig. 2. Example of ICU Finnish and its rough translation to English. Specific ICU Finnish features, such as spelling mistakes and other typographical issues, have been preserved in the translation. Due to patient privacy considerations, this text is an illustration preserving the ICU Finnish properties rather than an actual patient record.

In this paper, the linguistic analysis of ICU Finnish is from the sublanguage perspective. First, we present the key elements of the Harris theory of language. Then, we analyze the ICU Finnish in order to demonstrate how it differs from the standard Finnish and how a formal grammar for it should be constructed. While applying the framework of Harris in performing the sublanguage analysis, we will nevertheless also employ current mainstream linguistic terminology instead of only using the terms of Harris, relating the original terminology when necessary.

2.1 Departures from equiprobability and constraints on language

The starting point of Harris is that a theory of language must find the departures from equiprobability, i.e. to show that not all combinations of linguistic elements are equally likely to occur [14, pp.3]. As a trivial example, it is very unlikely that the sentence *Patient patient patient* exists. The distinction between the occurring combinations and the non-occurring ones can be made by defining constraints on the combinations taking place. A language defined by this constraint-seeking method thus consists of the word-sequences satisfying these combinatory constraints. In addition, its grammar reflects the information structure in the language. [14, pp.4-29]

Harris argues that the application of the constraint-seeking method also enables the organization of the information in a way that is specific to a given sublanguage. As a result, it is possible to compare different sublanguages by their specific constraints and information structures. [14, pp.18-23] These constraints imposed by the theory are termed dependence, inequalities of likelihood, paraphrastic reductions and linearizations. Together these elements define the conditions on word sequences that must be satisfied. In the following, we present these constraints.

Dependence Dependence forms the first constraint on which the other constraints rely. It defines the way in which the words in a sentence depend on others. In a sentence, words act as either operators or arguments. Operators, i.e. verbs, depend on arguments, that are e.g. nouns. Operators take different arguments according to the class they belong to [14, pp.53-61]. For instance, the operator *sleep*, which would traditionally be analyzed as an intransitive verb, takes just one argument, whereas the ditransitive *give* would take three. In other words, dependence deals with structures that would traditionally be seen as syntactic.

Inequalities of likelihood The dependence component does not impose any semantic restrictions; this is done by the inequalities of likelihood defining frequent and unusual combinations on operator-argument combinations. [14, pp.61-79] For instance, *patient* is reasonably likely to precede the operator *sleep*, whereas the argument *nurse* would have a lower likelihood with this operator. Friedman et al. [6] note that even though the constraints on combinations in general remain fuzzy, in sublanguages they can be very strict. For example, in science sublanguages, the possible combinations for the operator *attach* are more restricted than in general language.

Because of the different likelihoods on operator-argument combinations, the amount of information carried by different structures varies so that more frequent combinations carry less information than less frequent ones. A frequent combination may produce a low information situation, which produces omission, i.e. the zeroing of a very likely word. [14, pp.76-78] A common example of this phenomenon in our corpus is the zeroing of the subject which most likely refers to the patient and can therefore be safely omitted.

Reduction The third combinatory constraint, reduction, defines how complex sentences can be reduced to "base" sentences where the dependence and likelihood constraints are transparent. Similarly to the zeroing of very frequent components in a low information situation described above, also reductions occur when the linguistic component omitted carries little or no information; reductions only change the form of the sentence, not the information in it. [14, pp.79-96] For example, the sentence *The patient ate and slept* is reduced from *The patient ate and the patient slept.*

Linearization Finally, the last constraint, linearization, characterizes the possible sentences of the language together with dependence. Whereas the latter defines the possible operator-argument combinations, linearization takes place in the ordering of the combinations and also defines the respective order of operators and co-arguments. [14, pp.97-104] For example, the order of the arguments in the combination *The patient ate soup* and in the combination *Soup ate the patient* are realized by linear ordering.

2.2 Sublanguages

As we noted above, sublanguages differ from the general language by having different constraints and information structures. A more general way of defining them would be to depict them as subsystems of language that resemble largely the whole language but are restricted in some aspect. [11, pp.ix]

In practice sublanguages may also have more general properties of their own that the constraints of the general language would not allow [14, pp.272-273, 278-282]. For example, there are many standard Finnish constraints that the ICU Finnish does not fulfill, such as the obligatory use of a finite verb in a sentence. Because of these differences of constraints between sublanguages and the standard language, NLP applications made for a standard language are not necessarily applicable to sublanguages.

Constraints on sublanguages affect the language at all levels. In this paper, we are mostly interested in the ones having an effect on the construction of a parser for it. These include mostly lexical and syntactic issues, but also e.g. orthography is concerned, since frequent misspellings affect the parsing significantly.

In the next section, we will analyze these features in order to demonstrate how ICU Finnish as a sublanguage differs from the standard Finnish and provide crucial information for the construction of a formal grammar.

3 ICU Finnish as a sublanguage

ICU Finnish differs substantially from many of the sublanguages described in earlier studies which frequently focus on science sublanguages [6, 9]. Not being scientific text, having frequent misspellings, and allowing combinations constrained in standard Finnish, its features are notably different than those in scientific sublanguages with strict semantic, lexical, and stylistic constraints.

3.1 Syntactic features

The sublanguage features that perhaps most distinguish ICU Finnish from the standard language are syntactic. As is already shown by Friedman et al. [6] for clinical domain text in general, sentences are telegraphic, and often have zeroed elements that are clear in the context and thus reduced or zeroed according to the constraints. For example, the subjects of sentences are very often omitted. For instance, the subject of the sentence *the patient breaths weakly* carries in this context very little information and thus is often zeroed to *breaths weakly*.

Also arguments not referring to the patient can be clear from context. For example, procedures are often reported by verbs in passive voice so that the person performing the procedure remains unspecified (see Fig 3(a))[6]. Despite this, the agent is in practice apparent by the semantic constraints imposed by the inequalities of likelihood of the arguments and by the situational context.

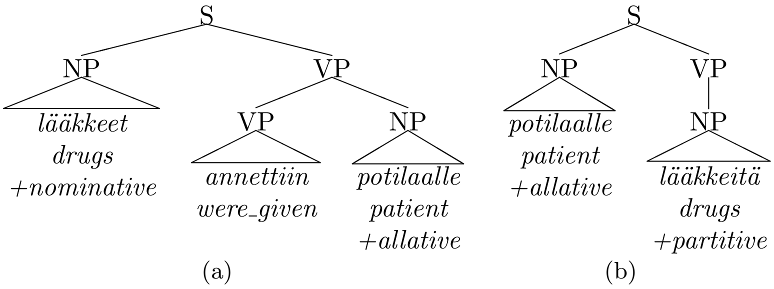


Fig. 3. Parse trees for (a) *drugs were given to the patient* and (b) *drugs to the patient*

[6] The examples are given in two forms and two rows. The first is the original Finnish example. The second lists the corresponding English words with the morphological cases specified when needed. The English translation for the examples are given in the caption.

A third subject omission type is the zeroing of anaphoric expressions having their referents in the preceding sentence. *Sister visited. [She] talked with the doctor.* Also in these sentences, the information carried by the omitted argument is so low that it does not need to be expressed at all.

In addition to subjects, also predicates, copulas and auxiliaries are often omitted in ICU Finnish, as is typical of telegraphic sublanguages [15]. This zeroing is also caused by a low information situation. Copulas, besides being frequent, carry little information by themselves and only have a syntactic function [16, pp.90,337]. For instance, there is no notable change in the information carried by *The patient is conscious* and *The patient conscious.*

Zeroed predicates occur often in sentences where the morphological case of one of the arguments defines the direction of the action and the way the arguments should be combined. The example in Figure 3(b) would be translated to English as *drugs to the patient*, the allative case in Finnish defining the direction of the action as *to the patient.* In standard Finnish, these clauses are traditionally considered as fragmentary [17, pp.839-840] despite the fact that they are used frequently in titles in for instance newspapers.

Another example of the zeroing of the predicate in a fragment where the case of the argument defines the direction is presented in Figure 4. The first example sentence in 4(a) represents the sentence without any zeroed elements. The first argument of this clause is a noun in the elative case (roughly corresponding to the English preposition *from*). The noun is followed by the operator *to come* [17, pp.860-861]. The reason for the predicate causing a low information situation is the morphological case of the first argument; elative being a semantic case[7], it defines the direction of the action making the operator secondary.

The information content of the predicate can be tested by omitting different elements of the sentence. In the sentence in Figure 4(a), the predicate is present. In the sentence 4(b), the predicate is zeroed with no significant loss in information. The example in Figure 4(c), with the subject omitted, leaves unspecified the starting point of the action thus omitting important information. The example in Figure 4(d), with the second argument missing, is ungrammatical with the operator *tulla* (Eng. *to come)* that would obligatorily need an argument in this context.

As a logical consequence of frequent subject and predicate omissions, the ICU Finnish sentences can consist of only noun phrases or even adjectival phrases (see Figure 5). Basically, these sentences have both the subject and the copula / auxiliary verb omitted. This feature is in fact very frequent in the ICU Finnish; 50 from the 80 sentences of our test treebank contain an omitted finite verb. This needs to be taken into account in the construction of a grammar, as will be shown in Section 5.

Finally, as a domain-specific sublanguage, ICU Finnish syntax has features that simplify the construction of a grammar. Most importantly, there are features of the standard Finnish syntax that do not take place in the sublanguage, the

[7] Cases displayed by adverbials are in Finnish called semantic cases; they take more concrete meaning than grammatical cases [17, pp.1175]

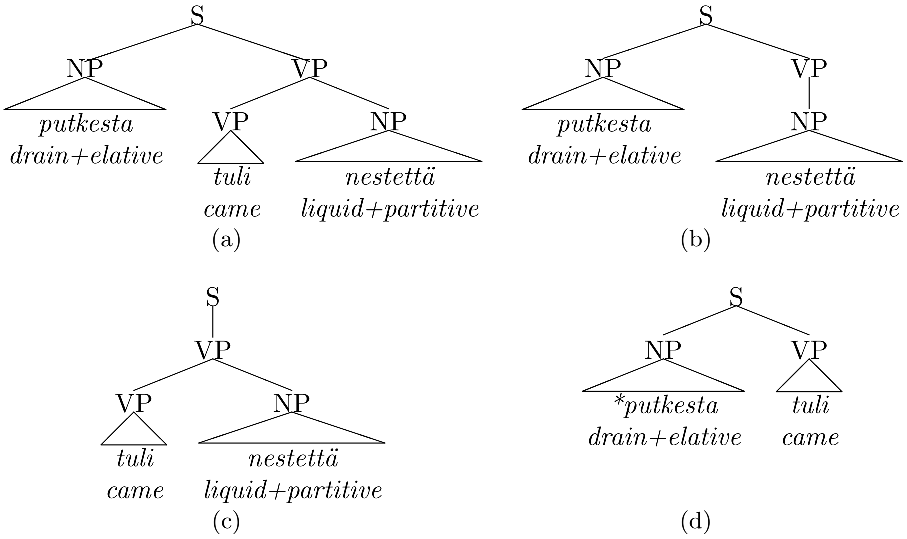


Fig. 4. Parse trees for (a) *liquid came out from the drain*, (b) *liquid from the drain*, (c) *liquid came out* and (d) **from the drain came*

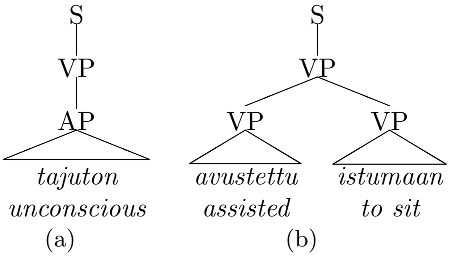


Fig. 5. Sentences with omitted subjects and (a) copula or (b) auxiliary

ICU Finnish consisting basically of reports about the treatment of the patient. For instance, as was illustrated in Figure 2, imperatives and interrogatives are extremely infrequent. Also interjections and discourse markers occur only rarely. Because of the infrequency of these features in ICU Finnish, they can be ignored in the development of the grammar making the development process simpler than it would be for the standard language.

3.2 Word order in ICU Finnish

In standard Finnish, the word order is most frequently S(ubject) V(erb) X(object or complement). However, because the roles of nouns can also be recognized by their case, the word order is not as strict as it would be e.g. in English. For

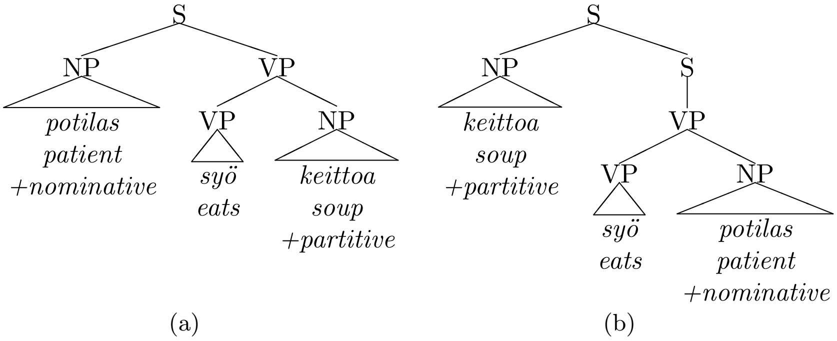


Fig. 6. Parse trees for (a) *patient eats soup* and (b) *it is soup the patient is eating*

instance, in Figure 6 *the patient* remains the subject and *the soup* the object of sentence even when they switch places.
Departures from the neutral SVX word order alter however the information structure of the sentence from neutral to marked [17, pp.1303]. For instance, in Figure 6(b), the fact that it is soup the patient is eating is stressed.

In fact, in ICU Finnish, linearization constraints allow word order change even when there is no apparent reason for it in the information structure. Thus they are less strict than those of standard Finnish. As another example, in an unmarked standard Finnish sentence, the modifier *a little* would generally follow the verb as it does in *liikehtii pikkuisen* (Eng. *moves a little*). In contrast, in ICU Finnish, its positioning in the beginning of the sentence, as in *pikkuisen liikehtii* (Eng. *a little moves*), does not necessarily modify the information structure.

These variations need to be considered in the development of a grammar for ICU Finnish, because a new rule must be made for every inversion of sentence elements (see Section 4.1).

3.3 ICU Finnish vocabulary

A restricted and domain-specific vocabulary is a typical feature of sublanguages [6]. ICU Finnish is no exception to this; its vocabulary differs substantially from that of standard Finnish. This affects the construction of a grammar for ICU Finnish, since FinTWOL is developed for standard Finnish and since its word recognition and morphological analysis are crucial in our system.

Words that are not covered by the FinTWOL lexicon are left with the label *unknown.* These are problematic since they cannot be assigned any lexical restrictions without special processing. This motivates efforts to resolve as many out-of-lexicon words as possible with the aim to assign them (at least approximate) lexical descriptions.

The principled way to deal with special domain words that are not found in the lexicon is to expand it to cover also the sublanguage vocabulary. A reasonable solution would have been to extend the lexicon coverage by vocabulary from

	Number of word types		
	ICU Finnish	Overlap of ICU and M. Rex	M. Rex
Total	57 709	3717	77 145
Covered by FinTWOL	36 785	3169	40 606
Covered by extended FinTWOL	40 260	3334	44 873
Not covered by extended FinTWOL	17 449	383	32 272

Table 1. Overlap between ICU Finnish and Metatesaurus Rex lexicons. Numbers indicate word types (unique inflected forms).

a domain lexicon such as the *Metatesaurus Rex*[8] which, among other vocabulary sources, includes FinMeSH, a Finnish translation of the MeSH thesaurus.[9] However, we found that the overlap between the ICU Finnish and Metatesaurus Rex lexicons is surprisingly small (see Table 1). Therefore, the lexicon of a version of the FinTWOL morphological analyzer was manually expanded by Lingsoft by approximately 3500 most frequent unknown word types in the ICU Finnish texts. These include the most frequent domain-specific terms such as *diureesi* (Eng. *diuresis*). The extension improved the number of running words recognized by FinTWOL by a relative gain of 42 %. While this manual effort considerably increased the FinTWOL coverage of ICU Finnish vocabulary, it did not remove the need for further unknown word processing and ICU vocabulary analysis.

To better understand the ICU Finnish vocabulary features, we analyzed a random sample of words for which the morphological analyzer failed to assign any description. For each such word, we determined first whether it is a (simple) variant of a known, in-lexicon word or out-of-lexicon, and then subcategorized these into commonly occurring cases. The results of this analysis are summarized in Table 2 and discussed in more detail in the following.

Abbreviations Abbreviations are very frequent in ICU Finnish; in contrast to standard Finnish, several abbreviations can even follow each other in a sentence as is shown in Figure 2. Most frequent among out-of-lexicon abbreviations are acronyms, such as *ICP* for *intra-cranial pressure*, which are often capitalized. Other cases include short forms of chemicals such as *CO2* for *carbon dioxide* and abbreviations derived from Latin (e.g. *po* for *per os*, meaning orally).

Abbreviations that are variants of in-lexicon words are typically truncated forms of full words such as *vas* and *oik* for forms of *vasen* and *oikea* (Eng. *left* and *right*, resp.). The truncated forms are not produced using standard Finnish rules: for example, *hemodynamiikka* (Eng. *hemodynamics*) is variously abbreviated at least as *hemodynam*, *hemodyn*, *hemod*, and *hemo*. ICU Finnish constraints on abbreviation formation thus seem less strict than in standard Finnish.

[8] www.terveysportti.fi/pls/rex/rex.metatesaurus.koti

[9] www.nlm.nih.gov/mesh

Variant of in-lexicon word	**54%**
Abbreviation	*21%*
Spelling variant	*20%*
Catenated number&unit	*7%*
Capitalization variant	*6%*
Out-of-lexicon	**42%**
Abbreviation	*31%*
Full word	*11%*
Other	**4%**
Split word	*2%*
Joined words	*2%*

Table 2. Frequencies of different types of unknown words. Major categories shown in bold, minor in italics. Relative frequencies shown separately for the two.

Spelling and capitalization variants Spelling variants are common in ICU Finnish; they occur notably more often that would be allowed in standard Finnish, likely owing in part to the time pressures of the environment. These can be divided into intentional variants and misspellings, misspellings being slightly more common. No domain-specific characteristics of typical misspellings were detected. However, many domain terms have regularly occurring variants that can be recognized as acceptable sublanguage usage: for example, the drug *noradrenaliini* (Eng. *noradrenaline*) is frequently alternately spelled as *noradrenalina* or *noradrenalin* and the breathing machine *Bennett* as *Bennet*. Additionally, ICU Finnish contains occasional instances of colloquial forms common in spoken Finnish, such as *tarvii* instead of *tarvitsee* (Eng. *needs*).

Further, capitalization is used relatively freely in ICU Finnish, with the most common departure from standard Finnish being lack of initial capitals for proper nouns such as drug names. Another frequent case is the use of fully capitalized words to mark section headings or emphasis.

Full words missing from the lexicon Approximately 10% of the remaining unknown words are “full” words, as opposed to e.g. abbreviations, that are not found in the FinTWOL lexicon. Roughly half of these are colloquial or informal words such as *labra* for *laboratorio* (Eng. *laboratory*), the other half representing mainly domain terms such as drug names and analysis methods or tools[10].

4 Implementation

In this section, we describe the technical aspects of the grammar implementation: the parsing framework, lexicon generation from FinTWOL output, out-of-lexicon word handling, and tokenization.

[10] A single general Finnish word, *huuliltaluku* (Eng. *lip reading*), was found in the analysis to be absent from the FinTWOL lexicon.

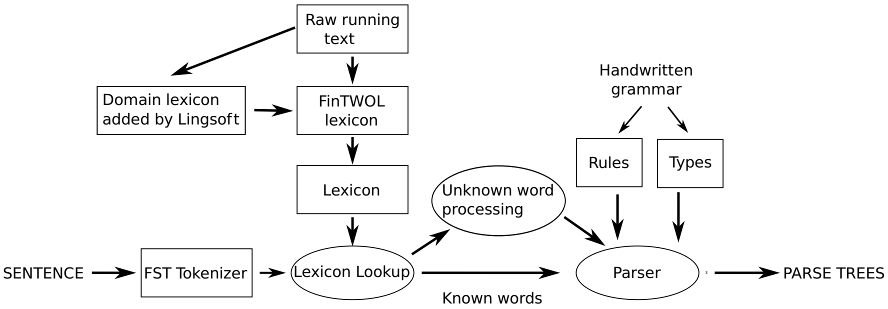


Fig. 7. Structure of the parser system

4.1 LKB

We develop the grammar in LKB, a lexicon and grammar development environment for formalisms based on typed feature structures (for detailed introduction to both LKB and typed feature structure grammars see [18]). The complete grammar is composed of three logical parts: the lexicon, the type system, and a set of grammar rules. We now describe these three components in greater detail, with a particular focus on the influence of the FinTWOL-based lexicon on the design of the grammar. The structure of the system is shown in Figure 7.

The lexicon of the grammar is written so that strictly no lexical information other than that provided by FinTWOL can be used. This allows fully automated generation of the lexicon based on FinTWOL analysis of text and avoids the need to develop a specialized domain lexicon. Consequently, the lexicon representation as typed feature structures has a one-to-one correspondence with FinTWOL output. This design principle comes at the cost of not being able to take a full advantage of a well-defined formal deep-syntactic formalism, such as HPSG [19] or LFG [20], as these highly lexical formalisms build on top of a rich lexicon encoding complex characteristics of the lexical entries not provided in the FinTWOL analysis.

An example of lexical entries generated from FinTWOL output is given in Figure 8. Note the special value *none* which is assigned whenever FinTWOL does not generate the corresponding tag. The interpretation of *none* is encoded in the grammar rules. For instance, when *inf* is *none* for a verb, the verb is in a finite form. In contrast, when *inf* is *none* for an adjective, it simply states that finiteness is not a feature that adjectives carry.

LKB grammar rules are defined as feature structures and parsing is performed via feature structure unification. In each step, feature structures representing the sub-constituents of a larger constituent are unified with the feature structure representing an individual grammar rule, resulting in a feature structure describing the newly formed constituent. A completed parse is thus represented as a complex feature structure obtained by unification of the feature structures corresponding to the lexical entries and the feature structures corresponding to

$$
\begin{array}{ccc}
\texttt{potilaan+N+GEN+SG} & \texttt{hengittää+V+PRES+ACT+SG3} & \texttt{runsain+A+SUP+NOM+SG} \\
\begin{bmatrix} \text{categ} & \text{Noun} \\ \text{orth} & \text{potilaan} \\ \text{base} & \text{potilas} \\ \text{mor} & \begin{bmatrix} \text{num} & \text{SG} \\ \text{case} & \text{GEN} \\ \text{voice} & \text{none} \\ \text{per} & \text{none} \\ \text{inf} & \text{none} \\ \text{deg} & \text{none} \\ \text{tense} & \text{none} \end{bmatrix} \end{bmatrix}
&
\begin{bmatrix} \text{categ} & \text{Verb} \\ \text{orth} & \text{hengitt} \\ \text{base} & \text{hengitt} \\ \text{mor} & \begin{bmatrix} \text{voice} & \text{ACT} \\ \text{per} & \text{SG3} \\ \text{tense} & \text{PRES} \\ \text{inf} & \text{none} \\ \text{num} & \text{none} \\ \text{case} & \text{none} \\ \text{deg} & \text{none} \end{bmatrix} \end{bmatrix}
&
\begin{bmatrix} \text{categ} & \text{Adj} \\ \text{orth} & \text{runsain} \\ \text{base} & \text{runsas} \\ \text{mor} & \begin{bmatrix} \text{num} & \text{SG} \\ \text{case} & \text{NOM} \\ \text{deg} & \text{SUP} \\ \text{voice} & \text{none} \\ \text{per} & \text{none} \\ \text{inf} & \text{none} \\ \text{tense} & \text{none} \end{bmatrix} \end{bmatrix}
\end{array}
$$

Fig. 8. Above: FinTWOL analysis given to the singular genitive noun *potilaan* (*patient's*), the active 3rd person singular verb *hengittää* (*[is] breathing*), and the singular nominative superlative adjective *runsain* (*[most] abundant*). Below: The corresponding feature structures generated from the FinTWOL analysis (slightly simplified for presentation). The main part-of-speech category is given by the feature *categ*, the surface wordform by *orth*, the lemma by *base*, and the morphological information is encoded in *mor*.

the grammar rules. This feature structure can be interpreted as a parse tree with nodes corresponding to the feature structures defining the constituents. The process is illustrated in Figure 9. The ICU Finnish grammar has 59 rules.

As a specific feature, each of the rules in the ICU Finnish grammar specifies its syntactic head by selecting the head word of one of its constituents. For example, the head word of a finite verb phrase is the head of the `Verb` nonterminal. The heads allow important information, such as noun cases, to be easily propagated upwards through the parse tree (see Figure 10). This information can then be constrained upon by further grammar rules.

The hierarchical type system, the third component of a LKB grammar, provides the ability to generalize the properties of grammar rules and lexical entries. As an example, we define the type *subj-case* to be either *nominative* or *partitive* and constrain the subject `NP` in a `NP VP` clause to be headed by a word whose case unifies with *subj-case* (as illustrated in Figure 9). The types are thus a tool for structuring the grammar along linguistic generalizations, decreasing the need for rule re-duplication in the grammar.

4.2 Unknown word processing

Unknown word processing is implemented as a separate step between morphological processing and parsing, and its implementation is motivated by the analysis of unknown word characteristics presented in Section 3.3. Unknown words are resolved in a cascade where the first rule that matches a word determines the lexical requirements assigned to it, and the rules are ordered roughly so that

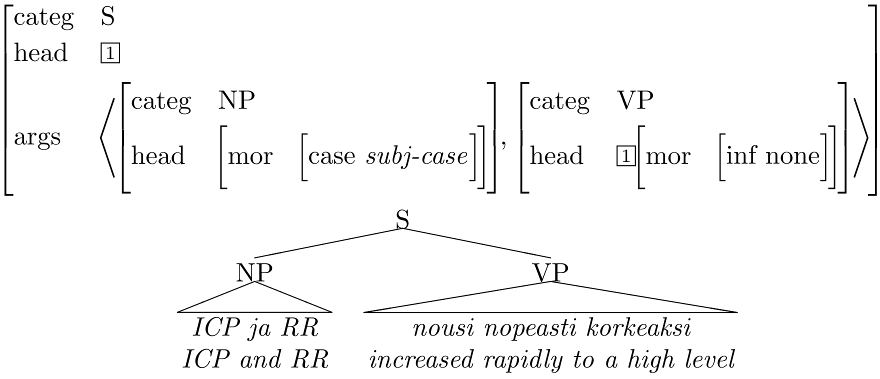


Fig. 9. The representation of the grammar rule $S \rightarrow NP\ VP$ as a feature structure (above) and an example application of the rule (below). The rule states that a noun phrase whose head is in a subject case (defined in the grammar) can combine with a verb phrase whose head is a finite verb. The head of the verb phrase becomes also the head of the sentence.

more specific, higher-confidence rules occur earlier and more generic rules later. The sequence of steps and the fraction of unknown words that they apply to are shown in Table 3.

Punctuation	15.1%
Numbers and numbers with units	2.6%
Known abbreviations	20.3%
Known colloquial words	0.1%
Explicitly marked case ending	3.0%
Capitalization variants	0.2%
Spelling variants	14.1%
All-capitals words	20.5%
Hyphenated compounds with in-lexicon word	2.6%
Possible case endings	10.0%
Unresolved	11.7%

Table 3. Fraction of unknown words captured by different steps of the unknown word resolution cascade.

Unknown words are resolved using a combination of surface feature-based heuristics, lexicon-based detection of variants of known words, and dictionary-based lookup of manually resolved sublanguage words that lack distinctive surface features.

Acronyms and many abbreviated forms (e.g. *CO2*) are recognized through surface features such as capitalization. A brief analysis indicated that when occurring in plain form they can typically be parsed as nominative nouns. Case endings are often explicitly marked for acronyms—for example, in *CPAP:lle* (Eng. *to CPAP*) the ending *lle* is written to mark allative case. All recognized acronyms are assigned noun lexical descriptions with the appropriate case if marked and nominative otherwise. The assignment of accurate lexical constraints to truncated abbreviated forms that have neither recognizable surface features nor explicit case markings is a difficult problem to which we have no computational solution; instead, we examined instances of the abbreviation in the corpus and determined which full forms would be appropriate in context. An expansion dictionary was compiled based on this analysis that is used to assign the lexical requirements of all of the full forms to each occurrence of an abbreviated form.

To detect and resolve misspellings as well as spelling variants of in-lexicon words, we implemented a simple spelling error correction algorithm that tests for deletions, insertions, replacements and transpositions of characters. Following preliminary experiments, only single-character modifications (edit distance one) are tested, and only words longer than four characters undergo spell-checking. Unknown words for which in-lexicon edit distance one variants are found are assigned the combined lexical descriptions of each of the variants: for example, the non-word *korkat* would receive the requirements of the words *korkeat* (Eng. *high*+pl) and *korvat* (Eng. *ears*). Capitalization variants are resolved by case-insensitive comparison of unknown words to in-lexicon words.

Partial lexical constraints for some unknown words are assigned based on case endings recognized by suffix matching. For example, the ending *-lle* relatively reliably indicates a nominal in allative case among out-of-lexicon words. Other endings are ambiguous: e.g. the ending *-sta* may indicate either an elative (e.g. *tuubista*, Eng. *[from the] tube*) or a partitive (e.g. *furesista*, Eng. *[some] Furesis*).

The development of the unknown word processing cascade was performed on a separate sample of the corpus that excluded annotated sentences to prevent possible overfitting of the gold standard data used for evaluation.

4.3 Tokenization and parsing

As shown in Figure 2, ICU Finnish text has very specific typographic properties such as frequent omissions of the space character as well as specific symbols (e.g. `-->`, +/-, etc.). We have therefore developed a special tokenizer for this text using the *xfst* finite-state tool [21] and use this tokenizer prior to parsing. The parsing itself is performed using the built-in parser in LKB with maximal chart size set to 15000 edges.

5 Evaluation

To support grammar development as well as measure the performance of the parser, we have developed a small gold-standard treebank of 80 sentences randomly selected from a large corpus of nursing notes. These 80 sentences comprise 614 tokens of which 489 are non-punctuation. For each sentence, we annotated

S-seq	sentences parsed [%]	sentences with fully correct parse [%]/[%]	best-parse F-score [%]	best-parse POS accuracy [%]	number of parses per sentence
S	56.2	55.6/31.2	91.0	93.4	11.0
SS	67.2	58.1/39.1	92.9	93.8	36.2
SSS	67.2	58.1/39.1	92.9	93.8	71.9
SSSS	65.6	59.5/39.1	92.9	93.9	72.2

Table 4. Results of the parser evaluation on the blind part of the ICU Finnish treebank. The *S-seq* column denotes the maximal number of full clauses allowed by the grammar under the top-level *S* nonterminal. The percentage of sentences for which the parser was able to produce a fully correct parse is reported as the percentage of all sentences that received a parse / the percentage of all sentences in the treebank. The POS accuracy is the percentage of POS tags correctly assigned by the parser.

the complete parse tree, assigning part-of-speech (POS) tags and non-terminal labels. Of the 80 sentences, 64 were not used in the grammar development and can thus serve as a test set for a small-scale evaluation of the parser.

We evaluate the parser performance using the standard labeled precision and recall as defined in the PARSEVAL measures [22] and implemented by the *evalb*[11] program; punctuation is disregarded. As the parser currently lacks a statistical parse-ranking component, we report the oracle best-parse performance together with the number of generated parses per sentence.

As can be seen from the Table 4, the number of parses expands significantly when a rule of the form `S`→`S S` is included. This is done to account for implicit coordinations, i.e. coordinations without an explicit surface marker, such as in *sisko soittanut potilas hereillä* (Eng. *sister called patient awake*).The addition of this rule, however, causes a combinatorial increase in the number of parses, especially due to the grammar allowing standalone nominal and adjectival phrases to form a full clause in order to deal with zeroed copula as discussed in Section 3.1. For example, adding the `S`→`S S` results in triplication of the average number of parses per sentence. In addition, the rules allowing a sequence of implicitly coordinated clauses substantially increase the size of the parse chart, even for sentences without implicit clause coordination. This can, paradoxically, lead to an effect where adding a rule into the grammar results in a decrease of the percentage of sentences that receive a parse, because the parse chart is more likely to exceed the maximal allowed number of edges. This effect is demonstrated in Table 4 for the rule allowing up to four implicitly coordinated clauses. The problem of implicit coordinations will require further systematic study in order to identify features that would help to constrain the parsing process and diminish the number of parses generated for these structures.

Depending on the treatment of implicit coordinations, the parser can currently analyze up to 67% of ICU Finnish sentences with a highly competitive best-parse F-score above 90%.

[11] http://nlp.cs.nyu.edu/evalb/

6 Conclusions and Future work

In this paper, we have presented a unification parser for ICU Finnish, the language used in daily nursing notes in a Finnish intensive care unit. We started by analyzing ICU Finnish in order to find out the features specific to this sublanguage. Then, we described the technical implementation of the grammar in the LKB system and presented a small-scale evaluation of the parser.

The analysis of ICU Finnish revealed that it has many features specific to clinical sublanguages; it is telegraphic with frequently omitted elements. A distinctive feature of ICU Finnish as a highly inflective language is the omission of a predicate in sentences where the case of the first argument specifies the direction of the action so that the predicate does not need to be expressed at all. Also the ICU Finnish vocabulary differs from standard Finnish by including domain-specific terms and frequent abbreviations. Overall, compared to standard Finnish, the syntax and lexicon of ICU Finnish are limited, which makes resource-efficient, rapid parser development possible.

In order to use the existing resources for Finnish, the lexicon is based on the output of the FinTWOL morphological analyzer. While this approach restricts the information available in the lexicon to that produced by FinTWOL, it is a crucial element in the construction of the parser. To reduce the negative effect of the words outside the FinTWOL lexicon on the parser performance, we developed an unknown word processing system resolving e.g. misspellings and abbreviations. The evaluation shows that even though further development is still needed, the construction of a parser by this method is feasible and resource-efficient. This method could also be applied to the resource-efficient construction of parsers for other sublanguages for which a morphological analyzer exists.

Apart from further increasing the coverage of the parser, there are two main directions to extend the current work. First is to build a parse-ranking model that will rank the ambiguous parses in order to identify the most likely parse. Several strategies for achieving this goal can be considered, including using the current parser for an efficient development of a corpus of sufficient size, or using the Finnish constraint grammar FinCG to gather lemma usage statistics, thus preventing parses with lemmas unlikely to occur in this domain.

Another direction for further research is to convert the parse trees into typed dependency trees, which have been suggested by many to be more suitable in applied setting (see e.g. [23]). As discussed in Section 4.1, each rule specifies its syntactic head. The resulting headed parse trees allow a straightforward construction of untyped, directed dependency trees. This is illustrated in Figure 10. Typed dependency trees could be derived from the untyped trees for example by the well-known methodology of [23]. An illustration of such a typed dependency tree is given in Figure 10. Additionally, the dependency scheme is likely to be a more suitable target for a careful evaluation of the parser performance (see e.g. Lin [24]) and for the possible development of a larger corpus.

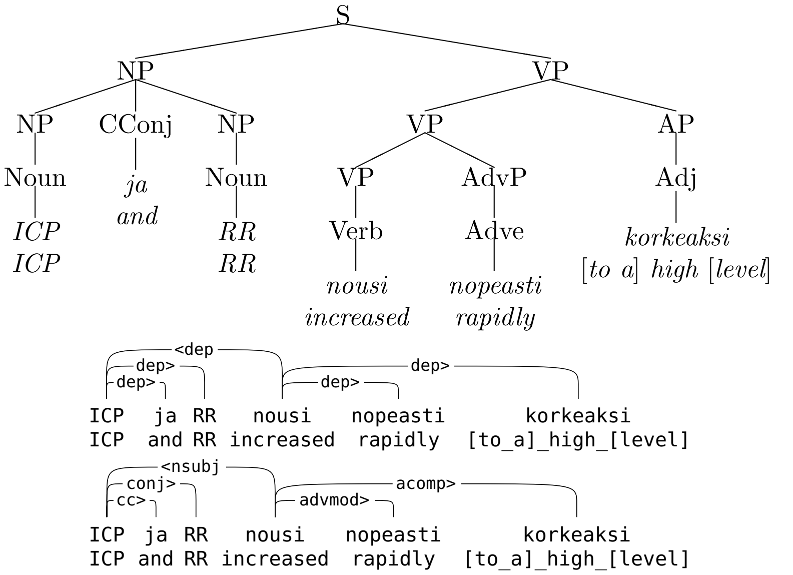

S
NP
VP
NP
CConj
NP
VP
AP
Noun
ja
and
Noun
VP
AdvP
Adj
ICP
ICP
RR
RR
Verb
Adve
korkeaksi
[to a] high [level]
nousi
increased
nopeasti
rapidly
<dep
dep>
dep>
dep>
dep>
ICP ja RR nousi nopeasti korkeaksi
ICP and RR increased rapidly [to_a]_high_[level]
<nsubj
conj>
cc>
acomp>
advmod>
ICP ja RR nousi nopeasti korkeaksi
ICP and RR increased rapidly [to_a]_high_[level]

Fig. 10. Top: The constituent tree with heads for the sentence from Figure 1. Middle: the untyped dependency tree automatically generated from the headed constituent tree. Bottom: the corresponding typed dependency tree in the Stanford scheme of de Marneffe [23].

Acknowledgments

This work was supported by the Academy of Finland and the Finnish funding agency for technology and innovation, Tekes. We are grateful to Simo Vihjanen and Sari Ahonen from Lingsoft Inc. for their cooperation on extending the FinTWOL lexicon.

References

1. Aronson, A.: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: Proceedings of AMIA 2001. (2001) 17–21
2. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics (2005) 173–180
3. Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., Salanterä, S.: Applying language technology to nursing documents: pros and cons with a focus on ethics. International Journal of Medical Informatics **76S2** (2007) 293–301
4. Sagae, K., Miyao, Y., Saetre, R., Tsujii, J.: Evaluating the effects of treebank size in a practical application for parsing. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing. (2008) 14–20

5. Koskenniemi, K.: Two-level model for morphological analysis. In: Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe, Germany, Morgan Kaufmann (1983) 683–685
6. Friedman, C., Kra, P., Rzhetsky, A.: Two biomedical sublanguages: a description based on the theories of zellig harris. Journal of Biomedical informatics **35**(4) (2002) 222–234
7. Friedman, C.: Sublanguage text processing — application to medical narrative. In Grishman, R., Kittredge, R., eds.: Analyzing Language in Restricted Domains. Lawrence Erlbaum, Hillsdale, NJ (1986) 85–102
8. Karvinen, K.: Hoitokertomusten kieli kotihoidossa. In: Kielitieteen päivät, Oulu, 2007. (2007) 129–130
9. Harris, Z.: Forms of information in science analysis of an immunology sublanguage. Dordrecht Kluwer (1989)
10. Kittredge, R., Lehrberger, J.: Sublanguage: studies of language in restricted semantic domains. de Gruyter (1982)
11. Grisham, R., Kittredge, R.: Analyzing language in restricted domains: sublanguage description and processing. Erlbaum (1986)
12. Aubin, S., Nazarenko, A., Nédellec, C.: Adapting a general parser to a sublanguage. In: Proceedings of RANLP'05. (2005) 89–93
13. Harris, Z.: Mathematical structures of language. New York Interscience Publishers (1968)
14. Harris, Z.: Theory of language and information a mathematical approach. Oxford Clarendon Press (1991)
15. Fitzpatrick, E., Bachenko, J., Hindle, D.: The status of telegraphic sublanguages. In Grishman, R., Kittredge, R., eds.: Analyzing Language in Restricted Domains. Lawrence Erlbaum, Hillsdale, NJ (1986) 39–52
16. Vilkuna, M.: Suomen lauseopin perusteet. Edita (1996)
17. Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T.R., Alho, I.: Iso suomen kielioppi / Grammar of Finnish. Suomalaisen kirjallisuuden seura (2004)
18. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI, Stanford, California (2002)
19. Pollard, C., Sag, I.A.: Head-Driven Phrase Structure Grammar. University of Chicago Press and CSLI Publications (1994)
20. Dalrymple, M., Kaplan, R.M., Maxwell, J.T.I., Zaenen, A., eds.: Formal Issues in Lexical-Functional Grammar. CSLI Publications (1995)
21. Beesley, K.B., Karttunen, L.: Finite State Morphology. CSLI Publications (2003)
22. Black, E., et al.: A procedure for quantitatively comparing the syntactic coverage of english grammars. In: Proceedings of the Fourth DARPA Speech and Natural Language Workshop, Pacific Grove, California. (1991) 306–312
23. de Marneffe, M.C., MacCartney, B., Manning, C.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC'06. (2006) 449–454
24. Lin, D.: A dependency-based method for evaluating broad-coverage parsers. Natural Language Engineering **4**(2) (1998) 97–114

Improving Inter-Rater Reliability Through Coding Scheme Reorganization: Managing Signs and Symptoms

Håkan Petersson[1], Hans Gill[1] and Hans Åhlfeldt[1]

[1]Department of Biomedical Engineering, Medical Informatics
SE-581 85 Linköping, Sweden
{hakan.petersson, hans.gill, hans.ahlfeldt}@imt.liu.se

Abstract. This study explores the potential for improving inter-rater reliability in diagnosis-based registries through the use of semantics. Cases originally coded as symptoms were reclassified based on location of manifestation, and inter-rater variability was measured through divergences of observed coding distributions from expected distributions and as variation in the degree of concentration across categories. Inter-rater variability decreased, however not statistically significant, and diagnostic categories with large variation in utilization rates were found. This calls for careful selection of topics for medical audit, knowledge discovery and other forms of information reuse. Although reclassification of symptoms may improve reliability, no straightforward association was found between diagnostic precision and contribution to overall variability. Nor could differences in diagnostic precision explain all variation within a diagnostic group. Further research on other dimensions of the coding system, as well as improved semantic models, are needed before symptom reclassification can be recommended as a general reliability-improving tool.

1 Introduction

The extensive use of computer-based medical record systems in general practice has established a foundation for reuse of information stored in clinical databases, such as in the evolving framework of quality assurance and medical audit [1]. One part of case documentation is usually the selection of a problem description in the form of a diagnostic label from a standardized coding system or controlled terminology. These entities constitute the patient's problem list and can be used in statistical compilations. Thus the diagnostic labels serve as entries to the record, problem descriptions in problem-based medical records, as well as data sources in aggregations for statistical purposes.

1.1 General Practice

In a review regarding medical decision-making in general practice, Brooke et al. assembled a set of features that distinguish between general practice and hospital practice [2]. With respect to conditions encountered, general practice deals with a wider range of problems and a greater mixture of psychosocial and pathophysiological problems. Diseases are encountered at earlier stages, and the focus is on ruling out serious diseases rather than formulating definitive diagnoses. Clinical problems in general practice have also been described as concerning the patients' total experience of 'illness' rather than a specific 'disease' [3].

Brooke et al. also conclude that the accepted degree of certainty is lower in general practice than in hospital practice [2], and Morrell presents a range for precise diagnoses of 30–90 percent, depending on the organ system in question [4]. Accordingly, diagnostic labels may include a large proportion of disease manifestation descriptions. Henceforth collectively referred to as symptoms, they consist of subjective indications of disease stated by the patient (symptoms) and objective findings observed by the physician (signs).

1.2 Reliability

Symptom assessment reliability in the sense of reproducibility has been assessed in different parts of the general practice spectrum, and studies include respiratory organs [5], the musculoskeletal system [6] and psychiatry [7]. Although the results of these studies are expressed in qualitative terms such as 'acceptable' or 'good' reliability, less degrees of inter-observer agreement on symptoms than on diseases has been reported [8].

Another complicating factor is the choice of a problem description representing a symptom or a disease that reflects the stage of the course of problem management at which the clinical concepts are sampled for the database, which is done by means of transformation into a standardized diagnostic description. This means that coders' preferences may lead to similar cases being labeled according to different diagnostic precision, and there are indications that reliability is low [9] and that the degree of clinical reasoning involved influences inter-rater reliability [10,11].

The credibility of annual activity compilations is constrained by the reliability of the underlying data, i.e. the diagnostic labels in the medical record. A recent, fairly large study confirmed the inherent difficulties of diagnostic coding others have reported before [12]. Furthermore, as clinical audit becomes a common approach to the improvement of clinical practice, the question of reliability must be considered in the choice of topics for audit, i.e. the diagnoses that are to be observed. Subjects—diagnoses or diagnostic groups—with high inter-rater reliability should preferably be selected.

1.3 Objectives

Since valid use and reuse of clinical data require high quality data on which subsequent processing can be based, quality is an important issue. Ideally, validity—the extent to which the coding represents the true status of the patients—should be measured. However, estimation of validity cannot be done without knowledge of the true status of each patient or a 'gold standard', which cannot be obtained. Reliability is a necessary but not sufficient condition of quality. If repeated measures are uniform, there is a chance that ratings are valid.

The aim of this paper is to study the potential for reducing inter-rater variability in general practice registries through the use of a semantic terminology model. By reviewing and comparing variation in data categorized according to different aggregation schemes, variability in different parts of a coding system can be analyzed [13]. For example, variation in diagnosis utilization rates, depending on diagnostic certainty, can be explored. Specifically, the role of diagnoses representing symptoms and findings will be examined in the context of the primary health care coding system.

2 Material and Methods

2.1 The Coding System

The coding system used in this study is KSH97-P, which is the official primary health care adaptation of the Swedish version of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10). The purpose of the modification has been to reflect common diagnoses based on sensible groupings [14]. Diagnoses correspond to either the three or the four character level in ICD-10, from which contiguous diagnoses have been aggregated into common joint diagnoses, and less frequent diagnoses have been added to residual groups. An example of the former is ‘Tinnitus’, while ‘Hearing loss, unspecified’ may serve as an example of the latter. This means that a diagnosis may refer to a single disease or symptom, as well as a group of diseases or symptoms. Furthermore, there are 30 diagnoses that refer to procedures such as ‘Supervision of pregnancy’ and ‘Contact for immunization against specified diseases’.

The main structure of KSH97-P is inherited from the ICD-10, i.e. a chapter structure reflecting the chosen principles of classification. For example, there is a chapter of perinatal diseases that are excluded by diseases related to organ systems. The classification uses 20 out of 21 ICD-10 chapters; the chapter of external causes is excluded due to the limited need for classification of injuries within primary health care [14].

The number of diagnoses in the version used is 972. Each diagnosis is described by a code, an original term and a preferred term. Original terms are taken from the Swedish translation of the ICD-10. Preferred terms, chosen by the Swedish Association of General Practice (SFAM), reflect the language used in medical records which means, for example, that coding technical tags such as NEC (Not Elsewhere Classified), NOS (Not Otherwise Specified), and Other are stripped. Sometimes alternative terms are synonyms to the preferred term and sometimes they represent a more specific concept. In addition, codes may have one or more alternative terms, which sometimes are synonyms to the preferred term and sometimes represent a closely related concept [11]. Altogether there are 1,566 alternative terms.

A state-of-health model has been implemented as a supplement and basis for the dynamic creation of subclasses according to location, etiology, and type of diagnosis [15]. Each diagnosis of the coding system is described by a category on each axis of the model. Table 1 contains the categories of the model. In addition, 30 diagnoses are described as procedures.

2.2 The Database

This study is based on a clinical database covering all patient visits to a particular primary health care unit during the three-year period (07/1997 to 06/2000). The total number of visits was 57,218. In 290 cases a diagnostic code was missing, and in 39 cases an invalid code was reported. Thus 99.5 percent of all cases were described by a valid diagnostic code.

The total number of coders was 66, out of which 47 had at least one case with a valid diagnostic code. Figure 1 illustrates the number of coded cases per coder, and as can be seen, visits are not evenly distributed among physicians. The reason for this is that some physicians have been employed longer than others during the time period studied. Based on their positions as general practice specialists or residents and the distribution of patient age, physicians were split into two homogenous groups. From each group, the eight coders with the largest number of coded cases were selected; in subsequent parts of the paper these will be referred to as senior and junior general practitioners (GPs). The groups accounted for 31,527 and 9,491 visits, respectively, and the number of cases per coder ranged from 2,076 to 5,604 in the former group and from 786 to 1,846 in the latter. Since illness patterns differ between men and women, the material was further divided with respect to patient gender.

Table 1. Dimensions and categories of the state-of-health model.

Location
Blood
Circulatory organs
Digestive organs
Ear
Eye
Inner secretory organs
Mammary gland
Multiple organs/functions
Musculoskeletal system
Nervous system
Psychological functions
Respiratory organs
Sexual organs
Skin
Urinary tract organs

Etiology
Circumstances
Deficiency
Endogenous
Infection
Injury
Mixed etiology
Poisoning

Type
Disease
Healthy
Risk
Symptom

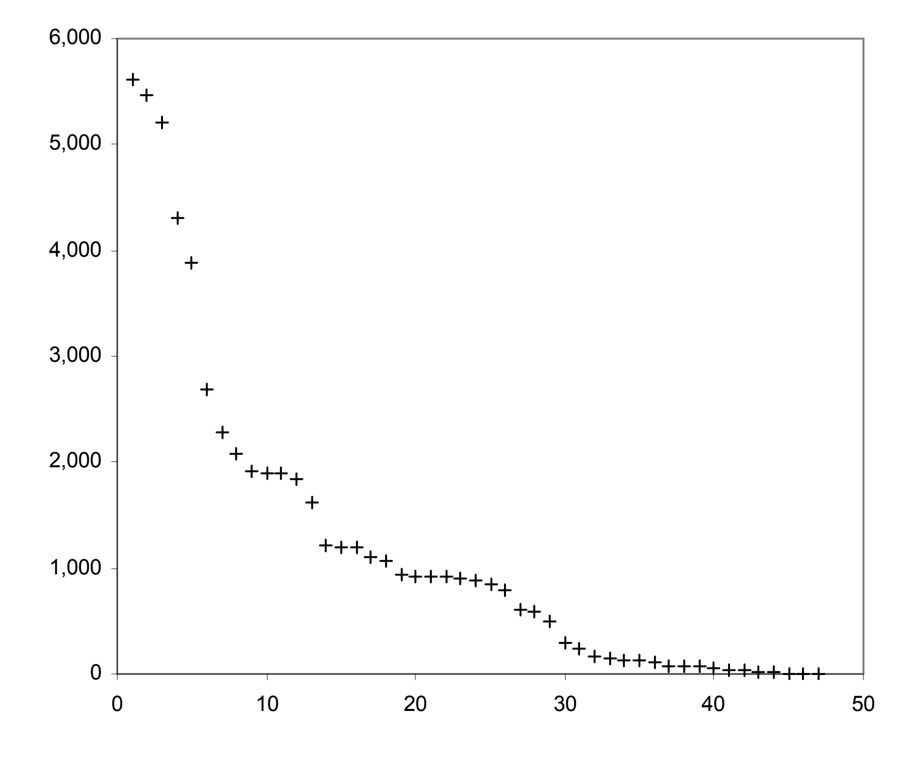


Fig. 1. The number of coded cases per physician.

2.3 Reclassification

Diagnoses classified as ‘Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified’ (Chapter 18) in the ICD-10 structure were examined regarding their state-of-health model descriptions. If they had a distinct location, i.e. a description other than ‘Multiple organs/functions’, they were reassigned to the corresponding organ system chapter. For example, the diagnosis Dyspnea (R06.0) was reclassified as ‘Diseases of the respiratory system’ (Chapter 10). Diagnoses described as procedures were reclassified as ‘Factors influencing health status and contact with health services’ (Chapter 21).

Coding scheme variation of the original classification and the reclassification were then compared. For the sake of comparison, the original structure was compared with the crude variation-reducing alternative of disregarding variation in Chapter 18 (hereafter referred to as the symptom chapter) by means of setting all coders’ frequencies to their expected frequency. Finally, the original chapter structure was compared with the combination of reclassification and disregarding the variability in remaining cases.

2.4 Variability

Inter-rater reliability was measured through the converse feature of inter-rater variability by means of the metric developed by Petersson et al. [13]. It measures divergences of observed coding distributions from expected distributions and is based on Levene's test for heterogeneity of variances with the median modification proposed by Brown and Forsythe [16]. Thus the variability associated with two coding schemes can be compared with one another. A p-value less than 0.05 was considered as statistically significant. The metric was also used for illustrating the categories' contribution to the total variation.

The metric involves a choice of something called alteration, which is a means of approaching the fact that coding scheme categories usually vary in utilization rate. We used the Shannon alteration (alteration C in the reference publication). This implies a weighing of absolute divergences by the amount of information received from the event of a code falling into the category. The measure of information quantity was used by Shannon [17] as the basis of the entropy of an information source. The standard deviation of the adjusted baseline divergences is then estimated as a measure of the coding scheme variability. In this calculation, the contribution of each category and coder to the total variability can be determined.

For each coding scheme, coder divergences were calculated with the material separated into four groups defined by the dichotomies 'Senior/Junior GP' and 'Female/Male patient.' The stratified data were then merged into four combined groups, one for each coding scheme, and the appropriateness of the pooling was verified by means of Kruskal-Wallis nonparametric analysis of variance (ANOVA).

2.5 Concentration

Concentration indices represent another means of comparing code utilization. The Herfindahl-Hirschman Index (HHI) of market concentration, used for example by the United States Department of Justice and the Federal Trade Commission in the analysis of proposed mergers [18], has been used in this context [19,20]. The HHI is calculated by summing the squares of each participant's percentage market share. Thus a market with only one supplier—a monopoly—is characterized by the maximum HHI value of 10,000.

In the context of variability, coders compare to markets, while coding categories substitute participants. The magnitude of the index is of little importance here; what matters is the variation among coders. Inter-coder consistency among physicians who treat comparable patient groups would yield similar HHI values. In this study, interquartile ranges were used as a measure of consistency.

It should be noted that there is no notion of order in the calculation of the HHI. If one coder uses the categories of a hypothetical coding scheme with the percentage distribution (10, 20, 30, 40) and another coder uses the same coding scheme with the distribution (40, 30, 20, 10), they will both have a HHI of 3000. Thus the concentration index will not indicate that—under the assumption that the coded populations are comparable—these two coders show a large degree of inter-coder variability.

It is also worth mentioning that concentration indices are by no means new in the field of reliability measurement. In the generalized unweighted kappa introduced by Fleiss [21], expected proportions are calculated as the summed square of proportion of cases falling into each category, which is interpreted as the probability that two independent observers agree on the classification of a particular case. Divided by the factor 10,000 then, the HHI of a market would give the probability that two independent customers use the same supplier.

3 Results

Out of all 41,018 assigned diagnoses, 12,540 (31 percent) are described as symptoms according to the state-of-health model. The symptom chapter contained 3,460 (senior GPs) and 1,548 (junior GPs) diagnoses, which means that of all diagnoses described as symptom in the state-of-health model, 40 percent were found in the symptom chapter.

Table 2 presents each chapter's proportional contribution to the total frequency and coding scheme variability. Junior physicians had a higher proportion of symptom chapter diagnoses—16 percent compared to 11 percent. This was their most frequently used chapter for female patients and second most frequently used chapter for male patients. Senior GPs had their fourth highest expected proportion of diagnoses in the symptom chapter for both female and male patients. With respect to inter-rater variation, the symptom chapter contributed less. With the exception of senior GPs' male patients, it had smaller relative variations than relative frequencies; the largest discrepancy was found for junior GPs' female patients. In both coder groups there was larger symptom chapter variation for male than for female patients.

Junior GPs had a higher proportion of symptom conditions with a distinct location—8.5 percent compared to 4.5 percent of the total number of cases. After reclassification, the number of cases in the symptom chapter was 2,057 for senior GPs and 743 for junior GPs. The number of diagnoses described as procedures was 16 in the senior group and 2 in the junior group.

Despite a tendency toward more variability among junior GPs, the Kruskal-Wallis ANOVA gave no evidence against the hypothesis that the pooled data were drawn from the same distribution (the four p-values ranged from 0.990 to 0.996).

Standard deviations for the coding schemes are presented in Table 3. Tests for heterogeneity of variances were performed on pooled data and compared with the original structure. Observed levels of significance (p-values) were 0.31 for the reclassified structure, 0.20 for disregarded symptom variation, and 0.055 for the reclassified structure with disregarded variation among remaining codes. Thus there were no statistically significant differences in inter-rater reliability between the coding schemes.

Results regarding variability in coding scheme concentration, presented in Figure 2, were inconsistent. While reclassification made junior GPs more homogenous in their category usage for male patient, the opposite was true for female patients. Senior GPs, who were more homogenous from start, became slightly less homogenous after reclassification but improved in the comparison structures.

Table 2. Coder- and gender-stratified percentage of frequency (Freq.) of coded cases and contribution to the total variation (Var.) per category. Gender refers to the gender of the patient. Due to rounding off, percentages do not total 100.

Coding system chapter		Senior GPs				Junior GPs			
		Female		Male		Female		Male	
		Freq.	Var.	Freq.	Var.	Freq.	Var.	Freq.	Var.
I	Certain infectious and parasitic diseases	2	1	2	1	4	9	5	16
II	Neoplasms	3	3	2	2	1	3	1	6
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	1	1	1	2	1	1	0	1
IV	Endocrine, nutritional and metabolic diseases	5	6	4	6	2	2	2	7
V	Mental and behavioral disorders	5	1	5	8	3	4	3	10
VI	Diseases of the nervous system	2	3	2	4	1	2	1	1
VII	Diseases of the eye and adnexa	2	1	2	1	2	2	2	3
VIII	Diseases of the ear and mastoid process	6	14	7	5	8	5	9	3
IX	Diseases of the circulatory system	11	5	14	21	6	6	8	4
X	Diseases of the respiratory system	15	16	18	6	17	14	17	6
XI	Diseases of the digestive system	4	6	4	3	2	2	3	1
XII	Diseases of the skin and subcutaneous tissue	4	1	5	3	6	7	6	10
XIII	Diseases of the musculoskeletal system and connective tissue	16	8	13	13	14	6	14	9
XIV	Diseases of the genitourinary system	9	20	4	4	8	23	4	5
XV	Pregnancy, childbirth and the puerperium	0	0	-	-	0	0	-	-
XVI	Certain conditions originating in the perinatal period	0	0	0	0	0	0	0	0
XVII	Congenital malformations, deformations and chromosomal abnormalities	0	0	0	0	0	0	0	0
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	11	7	10	13	17	5	15	9
XIX	Injury, and poisoning and certain other consequences of external causes	4	5	8	4	6	7	8	5
XXI	Factors influencing health status and contact with health services	1	2	1	5	2	3	2	5

Some discrepancies regarding relative frequency and contribution to coding scheme variation were found. For female patients, a high proportion of variation was observed in diagnoses classified as ‘Diseases of the genitourinary system’ (Chapter 14). With 9 and 8 percent of the total frequency, respectively, its part of the variation was 20 and 23 percent. On the other hand, ‘Diseases of the musculoskeletal system and connective tissue’ (Chapter 13) showed the inverse relationship with less relative frequency than variation.

For male patients, the largest inconsistencies were found in the chapter ‘Diseases of the respiratory system’ (Chapter 10), containing 18 and 17 percent of the cases, respectively, which represented 6 percent of the total variation for both senior and junior GPs. The junior GPs had a disproportionately large proportion of variation in ‘Certain infectious and parasitic diseases’ (Chapter 1) and ‘Mental and behavioral disorders’ (Chapter V).

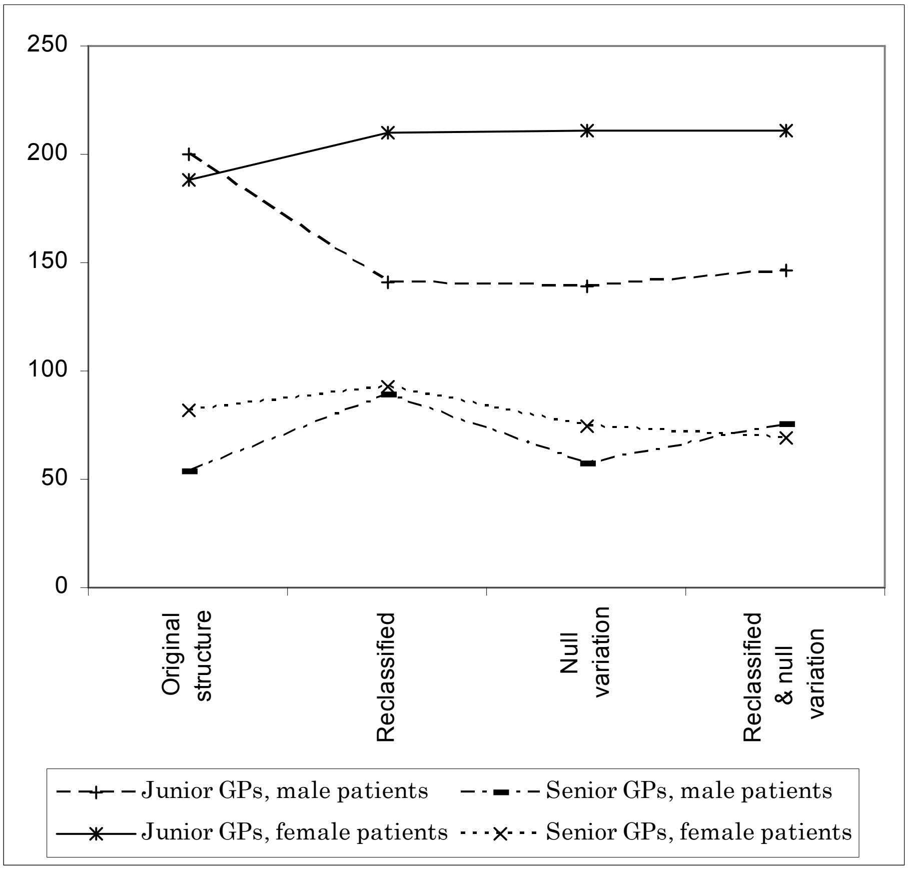


Fig. 2. Interquartile ranges for different study groups and aggregation schemes.

Table 3. Variability measured as standard deviation in four coding schemes: aggregation according to the chapter structure (Original structure), rearranged symptoms with distinct location (Reclassified), with disregarded variation in the symptom chapter (Null variation), and rearranged symptoms with disregarded variation in the remainder of the symptom chapter (Reclassified & null variation). Data are presented separately for senior and junior GPs as well as female and male patients, and for the four groups combined.

	Senior GPs		Junior GPs		Pooled
	Female	Male	Female	Male	data
Original structure	0.0159	0.0152	0.0177	0.0173	0.0165
Reclassified	0.0154	0.0148	0.0167	0.0166	0.0158
Null variation	0.0153	0.0140	0.0173	0.0163	0.0158
Reclassified & null variation	0.0148	0.0142	0.0166	0.0160	0.0154

4 Discussion

The ICD is a complex structure based on multi-dimensional categorization. While 60 percent of the diagnoses in the database described as Symptom in the state-of-health model were classified according to organ system of manifestation, etiology, and other circumstances, 40 percent were found in a separate chapter. One can argue about the implications of superseding the ICD-chapters by adding symptom diagnoses to organ system chapters. One line of reasoning is the appeal by Cimino to prevent semantic drift through rejection of the ‘Not Elsewhere Classified’ construct in coding system design [22]; the symptom chapter—named accordingly—is one such artifact, and reclassification may support comparability over time.

Statistical reporting also involves the issues of conformity to the aim of the analysis and the consistency with which the analysis is performed. If categories are created to reflect aspects of the practice we want to quantify, thus justifying a chapter of symptom diagnoses, this chapter should reasonably contain all symptoms—not only those that cannot be classified elsewhere. This calls for creation of new categories, independent of the ICD-structure. In the International Classification of Primary Care (ICPC), chapters are constituted according to organ systems and the subdivision into symptoms and diagnoses is done within each chapter [23]. Aggregation of ICPC-coded data is further discussed by Britt [24].

All coding scheme variations were reduced through the reclassification, but there was no statistically significant evidence that the inter-rater reliability was improved. Nor was there evidence that the hypothetical situation with no variability regarding the remaining diagnoses would improve overall reliability. One reason may be that there are too few symptom diagnoses to have any impact on the whole scheme. This suggestion is supported by the observation that compared to their senior colleagues, junior GPs had higher proportions of symptom chapter diagnoses that could be reclassified and slightly more effect from the reclassification (relatively larger reduction of the standard deviation).

The two features studied here—variability and concentration—represent attempts at quantifying reliability at different levels. While the variability metric suggested that reclassification may improve reliability on category level, the concentration comparison showed no support for improvement on coding scheme level. This sheds further light on the complexities of diagnostic coding and its associated issues.

The establishment of reference terminologies is indisputable. Large enterprises, such as SNOMED CT is on its way to widespread practical use, and the GALEN projects have turned into Open GALEN providing its Common Reference Model at no cost. Thus, a key element for providing semantic interoperability is within reach, offering terminological coverage of the entire clinical domain. How to best make use of such services to improve coding quality is an interesting research topic, but the results of this study indicate that reliability as a property of a coding scheme must be taken into account. Not only should terminologies used in annotation be examined to ensure valid end results, subsequent aggregation is also a process where different schemes may yield different results when e.g. used as training data in supervised learning.

Furthermore, compared to reference terminologies such as those mentioned above, the state-of-health model used here is a very coarse-grained structure. More detailed models would entail more flexible ways to aggregate data as soon as coded cases can be referenced to them. A reasonable assumption is that the more detail they provide, the more challenging the issue of reliability will be—not only due to the mere number of available options, but also because of inherent difficulties of the type investigated in this study.

The effect of the crude variation reduction was ambiguous. It gave the overall lowest standard deviation for senior GPs' male patients and the highest altered standard deviation for junior GPs' female patients. With respect to the pooled standard deviation, it had the same effect as the reclassification, but a lower observed level of significance. Nevertheless, reclassification must be considered the better alternative since it preserves diagnostic information.

A large contribution of a chapter to the total variation does not necessarily mean that large variation is to be found among categories within the chapter. The complement to over- or under-representation in an organ system chapter may be found in another chapter, such as the symptom chapter or the infectious disease chapter. Nor is intra-chapter variation necessarily reflected in inter-chapter variation—differences in coding preference, such as the choice of symptom or disease description, may be cancelled out within the chapter so that the coders have identical total observed relative frequencies.

The chapter 'Diseases of the genitourinary system' was further examined with respect to type and location of its diagnoses in the state-of-health model. For female patients, the proportion of cases described with the type category Symptom ranged from 23 to 53 percent (senior GPs) and from 5 to 36 percent (junior GPs). This spread was not larger than in chapters such as 'Diseases of the circulatory system' and 'Diseases of the respiratory system'. In the location dimension, 42 percent (senior GPs) and 36 percent (junior GPs) of the chapter variation concerned the group 'Multiple organs/functions', which only contains one used diagnosis: Menopausal and female climacteric states (N95.9P). Although this is the second most prevalent diagnosis of this chapter, the over-represented variability contribution implied low reliability of this particular diagnosis. Compared to this rather vague symptom cluster, the most common diagnosis, Cystitis, described as a disease, belongs to the less variable location category Urinary tract organs. It contained 42 and 63 percent of the cases, respectively, and explained 34 percent of the senior group variation and only 8 percent of the junior group variation.

Morrell reported a low degree of diagnostic certainty for conditions concerning the digestive system, and a high degree for conditions related to the skin and respiratory organs [4]. In our study, chapters associated with these organ systems have 21, 83, and 91 percent of the cases diagnosed as Disease, respectively, in the state-of-health model, which is in line with those findings. There was no indication of a relationship between the degree of diagnostic certainty and inter-rater variability for these chapters. In addition, for female patients, Chapter 14 had 65 percent disease diagnoses but nevertheless a large variation, and in Chapter 10 with relatively large differences in variance contribution among the four groups, the distribution of Symptom diagnoses was homogeneous.

A reason for the insignificant results may be that the metric does not have enough power to detect differences among the coders. Inter-rater variability due to other factors, such as differences in patient populations not taken into consideration, may obscure variability due to coding differences. Although no clear evidence was found in this setting, the observed levels of significance are low enough to motivate further analysis of systematic differences in the selection of problem descriptions.

5 Conclusion

The state-of-health model provided a means by which to analyze coded data independent of the ICD structure and to reorganize this structure into a more uniform view. Reclassification of symptom diagnoses showed a potential lowering of the inter-rater variability at the chapter level. However, neither a chapter's average proportion of symptom diagnoses, nor its inter-coder variation of symptom diagnoses seemed to be associated with the contribution to total variability. In addition, differences in diagnostic precision may not explain all variation within a chapter.

Since variation within categories and variation between categories are not necessarily associated, future research should include both. This also suggests that when aggregated data are compared, it is important to measure reliability at the proper level of abstraction, i.e. according to a particular aggregation scheme. While data may be reliable in one dimension of aggregation, it may not be in another dimension of aggregation.

Diagnoses and diagnostic categories with large variation in diagnosis utilization rates were found, and further analysis of these findings may provide insight into the mechanisms of inter-rater variability and how the reliability of clinical databases can be improved. This calls for care in the choice of audit topics, areas for knowledge discovery and further exploration of the relationship between diverging category proportions and inter-rater agreement.

Acknowledgments

This study was supported by grant 621-2002-4052 from VR (The Swedish Research Council). The authors would like to thank Lars Berg at the Centre for Epidemiology at the National Board of Health and Welfare for helpful comments, and Tibro Primary Health Care Center for providing the database. Approval for this study was obtained from the local ethics committee.

References

1. Månsson, J., Nilsson, G., Björkelund, C., Strender, L.E.: Collection and Retrieval of Structured Clinical Data from Electronic Patient Records in General Practice: A First-phase Study to Create a Health Care Database for Research and Quality Assessment. Scand. J. Prim. Health. Care 22, 6--10 (2004)
2. Brooke, J.B., Rector, A.L., Sheldon, M.G.: A Review of Studies of Decision-Making in General Practice. Med. Inform. (Lond) 9, 45--53 (1984)
3. Royal College of General Practitioners: The Future General Practitioner: Learning and Teaching. The British Medical Journal for the RCGP, London (1972)
4. Morrell, D.C.: Symptom Interpretation in General Practice. J. R. Coll. Gen. Pract. 22, 297--309 (1972)
5. Stavem, K., Jodalen, H.: Reliability and Validity of the COOP/WONCA Health Status Measure in Patients with Chronic Obstructive Pulmonary Disease. Qual. Life. Res. 11, 527--533 (2002)
6. Jinks, C., Lewis, M, Ong, B.N., Croft, P.: A Brief Screening Tool for Knee Pain in Primary Care. 1. Validity and Reliability. Rheumatology (Oxford) 40, 528--536 (2001)
7. Terluin, B, van Hout, H.P.J., van Marwijk, H.W.J., Adér, H.J., van der Meer, K., de Haan, M, et al.: Reliability and Validity of the Assessment of Depression in General Practice: The Short Depression Interview (SDI). Gen. Hosp. Psychiatry 24, 396--405 (2002)
8. Williams, H.C., Burney, P.G., Strachan, D., Hay, R.J.: The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. III. Observer Variation of Clinical Diagnosis and Signs of Atopic Dermatitis. Br. J. Dermatol. 131, 397--405 (1994)
9. Nilsson, G., Petersson, H., Åhlfeldt, H., Strender, L.-E.: Evaluation of Three Swedish ICD-10 Primary Care Versions: Reliability and Ease of Use in Diagnostic Coding. Methods Inf. Med. 39, 325--331 (2000)
10. Hagelin, E.M.: Coding Data from Child Health Records: The Relationship between Interrater Agreement and Interpretive Burden. J. Pediatr. Nurs. 14, 313--321 (1999)
11. Petersson, H., Nilsson, G., Strender, L.-E., Åhlfeldt, H.: The Connection between Terms Used in Medical Records and Coding System: A Study on Swedish Primary Health Care Data. Med. Inform. Internet Med. 26, 87--99 (2001)
12. Stausberg, J., Lehmann, N., Kaczmarek, D., Stein, M.: Reliability of Diagnoses Coding with ICD-10. Int. J. Med. Inform. 77, 50--57 (2008)
13. Petersson, H., Gill, H., Åhlfeldt, H.: A Variance-Based Measure of Inter-Rater Agreement in Medical Databases. J. Biomed. Inform. 35, 331--342 (2002)
14. National Board of Health and Welfare: Klassifikation av sjukdomar och hälsoproblem 1997. Primärvård [Classification of Diseases and Health Problems 1997. Primary Health Care, in Swedish]. Socialstyrelsen, Stockholm (1997)
15. Petersson, H., Nilsson, G., Åhlfeldt, H., Malmberg, B.-G., Wigertz, O.: Design and Implementation of a World Wide Web Accessible Database for the Swedish ICD-10 Primary Care Version Using a Concept System Approach. In: Masys, D.R. (ed) Proc. 1997 AMIA Annual Fall Symposium. p. 885. Hanley & Belfus, Inc, Piladelphia (1997)

16. Brown, M.B., Forsythe, A.B.: Robust Tests for the Equality of Variances. J. Am. Stat. Assoc. 69, 264--267 (1974)
17. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27, 379--423 (1948)
18. U.S. Department of Justice and the Federal Trade Commission: Horizontal Merger Guidelines Issued 1992, Revised 1997. Retrieved June 18, 2008, from http://www.usdoj.gov/atr/public/guidelines/hmg.pdf
19. Spangler, W.E., May, J.H., Strum, D.P., Vargas, L.G.: A Data Mining Approach to Characterizing Medical Code Usage Pattern. J. Med. Syst. 26, 255--275 (2002)
20. Lungeanu, D., Zaharie, D., Holban, S., Bernad, E., Bari, M., Noaghiu, R.: Exploratory Analysis of Medical Coding Practices: The Relevance of Reported Data Quality in Obstetrics-Gynaecology. Stud. Health. Technol. Inform. 136, 839--844 (2008)
21. Fleiss, J.L.: Measuring Nominal Scale Agreement Among Many Raters. Psychol. Bull. 76, 378--382 (1971)
22. Cimino, J.J.: Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Methods Inf. Med. 37, 394--403 (1998)
23. Lamberts, H., Wood, M. (eds): ICPC—International Classification of Primary Care. Oxford University Press, Oxford (1987)
24. Britt, H.: A New Coding Tool for Computerised Clinical Systems in Primary Care—ICPC Plus. Aust. Fam. Physician 26 Suppl 2, S79--S82 (1997)

Nurses Translating Technology

Helena Karsten[1,2], and Riikka Vuokko[1,2]

[1] Department of Information Technologies, Åbo Akademi University
[2] Turku Center for Computer Science (TUCS)
Joukahainengatan 3-5, 20520 Turku, Finland
{Eija.Karsten, Riikka.Vuokko}@abo.fi

Abstract. Nursing work is being transformed by the implementation of electronic patient record systems. The new information systems affect not only the documenting, collecting and sharing of patient information, but also the coordination of work and routine work practices. This paper presents a view of nurses translating new information technology to make it an inseparable part of their daily working. The emerging view on nursing work is based on analysis of a large amount of data, as the work transformation has been studied in the hospital already from the year 2001 onwards with observations and interviews. In the data, nursing work emerges as complex and intensive collaboration with other participants in care and with various technologies. The findings from the study show that there are considerable challenges to translate new technology in a complex environment. Challenges emerge from mundane issues, such as positioning of the information sources in a ward or in a clinic environment, visibility of information and task performance in the changing situation, interdependencies of documenting, and actual breakdowns of the new information system.

Keywords: Electronic patient record, technology adoption, work practices, Actor-network theory.

1 Introduction

While nursing work is slowly adjusted to increasing computerizing and record-keeping, the organizational implementation of information technology in hospitals is often presented as complex and problematic (e.g. Berg 2001, 2004; Jones, 2003; Moser & Law, 2006). For example, Scott et al (2005) describe a case where the implemented electronic medical record application was later withdrawn from use, due to users' resistance, reduced productivity, and technical problems. Organizational implementation situations have been characterised as phases of a political conflict (Latour, 2005; Woolgar, 1991), as a culture conflict (Leidner & Kayworth, 2006), or as a restructuration of technology-in-use (Orlikowski, 2002). We have studied work in a hospital since 2001, in eight different wards or clinics. In this paper, we explore how hospital staff – predominantly nurses – adopt and adapt new information technologies in the four cases selected.

In diffusion studies, adoption of a new information technology has been studied to find out facilitators or drivers behind individual users' adoption decisions (Rogers, 2003). The behaviour of the organizational members as adopters and users of technology have been studied with various features of, for example, environmental, organizational, individual, or technological characteristics (Gallivan, 2001; Jeyaraj et al. 2006; Wainwright & Waring, 2007), but no constant pattern has emerged. In this paper we propose that the adoption environment in a hospital context is a complex and dynamic one. Instead of individual adoption decisions or characteristics, we aim to understand adoption of new information technologies as co-constructed in multiple negotiations and intertwined relationships within the working environment.

Work in a hospital is situational and even unpredictable by nature, more than in office environments. Also, the focus of work is on the unexpected: sudden turns to worse in an illness, babies born at all hours, traffic accidents happening. At the same time, hospital environments are complex and difficult to study due to disciplinary boundaries and ethical requirements involved (Wainwright & Waring, 2007, Moser & Law, 2006). The work practices are based on shared sets of rules, such as the laws concerning public health services, the hospital standard procedures, or various health care standards and classifications. Instead of standardising work, these all seem to add up on the organizational complexity where many different services contribute to advance the patients' health. In this study, we explore work in hospitals as a socio-technical assemblage, where various actors form and re-form work practices and adapt to computerization in a jointly constructed manner (Latour, 2005; Jacucci et al., 2006).

Our main research focus has been on how new information systems are taken into use and how they have been gradually built into the hospital work practices. The organizational implementation has progressed rather slowly from one hospital unit to a next one, and from one application to a next one. The software provider has slipped on schedules and the parts acquired have not always met the specifications. This combined with technical (e.g. data transmission) and integration problems have caused a considerable amount of resistance amongst hospital staff, resulting in, for example, change of the project leader, and lengthening of total adoption time in the hospital.

The hospital in question is one of the major teaching hospitals in Finland, and it was one of the first ones to start organizational implementation of an integrated multimedia electronic patient record system. Developing the hospital information technology environment was started in the 1970s. Many applications were installed in the early 1980s and several of these are still in use. In total, there are circa 100 applications in the hospital, plus various units have their "own" applications. The hospital IT office is outsourced and the new company (as of 1.1.2008) has a staff of 66 persons who mainly do project management, training and technical support. All software is bought from private software houses. Integrating the various applications has been and continues to be the main challenge. On the national level, this has been also realised, with much standardisation already in place and a national archive building in progress.

2 Complexity and network of actants

Complexity of hospitals as working environments has been growing due to hospital operations and administration becoming increasingly dependent on information, communication and decision support technologies (Berg, 2004; Moser & Law, 2006; Wainwright & Waring, 2007). Due to increased connectivity to, for example, municipal health centres, private clinics, pharmacies, or government agencies, this complexity is expected to continue growing (Cohen, 1999; Merali, 2004; Jacucci et al, 2006). National policies may, as well, contribute to increased dynamism, uncertainty and discontinuity in the hospital life. While the electronic patient record (later ERP) systems have been introduced with a promise of increasing quality and traceability of health care services (e.g. Berg, 2004), organizational implementations of EPR systems have been met with all sorts of attitudes, issues and barriers (Berg 2001, 2004; Jones, 2003; Karsten & Laine 2007; Scott et al, 2005). Practical organizational issues range from re-organization of work coordination to questions of responsibility and roles in documenting patients' health care. Technical problems easily add to the frustration of the users, especially when reserve systems are not in place

There are several studies of adoption and decision making in strongly defined groups, such as medical and nursing professionals or researchers, using Actor-Network Theory, ANT (Latour, 2005, 1999, 1991). These have attempted to describe large and complex networks of technical innovation and change (Berg, 2001; Moser & Law 2006; Moser, 2005; Middleton & Brown, 2005; Latour, 1999). There has been considerable attention also on complexity issues (Moser & Law, 2006) increased in organizations by new connections and combinability between people, technology and information (Merali, 2004). In ANT, networks constitute of a relevant social group (Bijker, 1995) of actors that negotiate and interact with each other to solve a shared "problem" such as how to use unfamiliar technology in practice. In our study, the main social actors are nurses, physicians and ward secretaries. While not underestimating the importance of human actors, we include also other relevant actors or actants, such as technical artefacts such as vitality monitors, organizational rules and scripts, as well as patient records.

In ANT, new innovations are taken into use when a relatively stable heterogeneous network of aligned interest is created and maintained (Latour, 2005, 1999). In the hospital environment, different wards and care units can, for example, form such networks of aligned interests when shared objectives emerge or break the daily routines. Successful implementation and adoption of new technologies involves the building of alliances between various actors, and this includes individuals and groups, as well as other entities such as machines. Thus, both the social and the technical are involved as the actors are enrolled into a network. As the network evolves, the nature of the project and the identities and interests of the actors are themselves transformed through negotiations and sharing of information. The results of the transformation process, the translation, are subsequently inscribed into technologies, in organizational practices and routines (Latour, 2005).

According to Latour (2005, 1999) translation of technology into use involves careful fitting and negotiating. The process of translation is started when a problem of interpretation emerges about the nature and use of the new technology. When new technology is taken in to use and its utilization becomes a routine, the technology disappears form the core of consciousness (Ciborra, 2002; Orlikowski, 2002), as its use no longer demands special efforts and it is embedded in the context. Latour (1999) describes that the translation process can continue to a point where the new technology becomes embedded in its use context as a simplified black box. Black boxing occurs when, for example, the organizational actors can use a solution or a tool although they cannot tell its inner functions. In a way, here, a complicated technical artefact can be simplified as a black box that can be utilized but without special efforts by the users. In the hospital environment, the organizational implementation of the EPR system seems to be in various stages of translation and black boxing.

Network has been a prevailing metaphor for studies emphasising connectedness or organizational co-construction of new technology (Merali, 2004; Latour, 2005, 1999). A network is dynamic, and has flexibility and adaptability to survive. In networks, interconnectivity has been described as negotiable, as voluntary or open-ended, or even as unpredictable. As such, the metaphor has fitted well to describe contemporary organizations and the changes in working life. In research, it means recognition of fragmentation and complexity (Knox et al., 2006). Still, the network metaphor has been criticised to lack clear definitions, or to have multiple meanings (Cohen, 1999; Doolin and Lowe, 2002; Latour, 2005). There is no agreement of what kind of nodes and relations a network may consist of. As such, power relations can be left undefined or even neglected when using the network metaphor. In this study, the actors in of the hospital units are considered as “the nodes” of their working network, and no individual actor or a stakeholder party is approached as a sole controller of complex interaction.

3 The four cases

Our research program began in 2001 when the hospital set up the project to carry out the organizational implementation of the integrated electronic patient record (EPR) system over the next five years. We were able to follow the implementation for the duration of the project. The implementation was to be carried out in steps, and each hospital unit was free to decide when and what parts of the system they would use. MD-Oberon for controlling and administrating health care processes was taken into use between the years 2000-2002. MD-Miranda, the main part of the EPR, was built to work together with it. In studying Miranda use, we have focused on the nursing records with parts for care planning, care giving, follow-up and evaluation. In special areas of care, such as in a delivery ward or in intensive care, there are also unit specific systems for care.

The process of organizational implementation and adoption of information technology have been explored with qualitative methods (e.g. Taylor & Bogdan 1998). The observations in various hospital units are combined with semi-structured interviews with medical staff, including surgeons, physicians and nurses, as well as representatives from the hospital administration, such as unit secretaries, and participants of the EPR project team. All of the interviews were tape-recorded and later transcribed. Observations were documented in field notes and these were supplemented with photographs of technical artefacts and videos of their use.

Understanding hospital work practices means understanding wider cultural phenomena in a given context. In this sense, shared practices are ways to distinguish oneself from others, from outsiders. Still, work practices are hard to observe as they are internal to individuals and, as such, often invisible to an outside observer. While practices may form an invisible, taken-for-granted set of attitudes and reward structures (Haythornthwaite, 2006), in a sense, these practices contribute to a transparent infrastructure that is inherent to a community (Star and Ruhleder, 1996). According to Haythornthwaite (2006), practices are instantiated in the technologies used to accomplish work, and as such, we aim to observe work practices as manifested by the use of information technology. Next, the four cases are introduced with an emphasis on the practice level of accomplishing daily work. These four units are an oncology ward, a surgical outpatient clinic, a maternity ward, and a paediatric intensive care unit.

3.1 Oncology ward

Our research program began in 2001 when the hospital set up the project to carry out the organizational implementation of the integrated electronic patient record (EPR) system over the next five years. We were able to follow the implementation for the duration of the project. The implementation was to be carried out in steps, and each hospital unit was free to decide when and what parts of the system they would use. MD-Oberon for controlling and administrating health care processes was taken into use between the years 2000-2002. MD-Miranda, the main part of the EPR, was built to work together with it. In studying Miranda use, we have focused on the nursing records with parts for care planning, care giving, follow-up and evaluation. In special areas of care, such as in a delivery ward or in intensive care, there are also unit specific systems for care.

The process of organizational implementation and adoption of information technology have been explored with qualitative methods (e.g. Taylor & Bogdan 1998). The observations in various hospital units are combined with semi-structured interviews with medical staff, including surgeons, physicians and nurses, as well as representatives from the hospital administration, such as unit secretaries, and participants of the EPR project team. All of the interviews were tape-recorded and

later transcribed. Observations were documented in field notes and these were supplemented with photographs of technical artefacts and videos of their use.

Understanding hospital work practices means understanding wider cultural phenomena in a given context. In this sense, shared practices are ways to distinguish oneself from others, from outsiders. Still, work practices are hard to observe as they are internal to individuals and, as such, often invisible to an outside observer. While practices may form an invisible, taken-for-granted set of attitudes and reward structures (Haythornthwaite, 2006), in a sense, these practices contribute to a transparent infrastructure that is inherent to a community (Star and Ruhleder, 1996). According to Haythornthwaite (2006), practices are instantiated in the technologies used to accomplish work, and as such, we aim to observe work practices as

manifested by the use of information technology. Next, the four cases are introduced with an emphasis on the practice level of accomplishing daily work. These four units are an oncology ward, a surgical outpatient clinic, a maternity ward, and a paediatric intensive care unit.

3.2 Surgical clinic

The surgical clinic was first visited during 2001-2 and then, with the major crisis described here, again during 2006. The data from 2001-2 consists of eight interviews and 18 hours of observation. The outpatient clinic works in a very rapid pace. A surgeon and a nurse meet a patient for a discussion, and subsequent measures are agreed on. The nurse prepares for this meeting by gathering all data of the patient, including the paper-based patient record. Surgeons in ten different specialities come from their wards to meet the patients. Thus scheduling is tight and preparations are started many days ahead. The day surgery has a slower pace. A team of specialists operates on a patient that is usually otherwise healthy.

Ten staff members from the surgical clinic and three EPR project members were interviewed during spring 2006. To complement the view from the interviews, there was also a video recording of an actual use situation, where a surgeon and a nurse care for one of the in-bed patients and document the care on both paper and in the EPR.

In 2001-2 the surgeons had been interested in the new software and eager to implement it. With the problems of the technical facilities and the software, this soon turned into disillusionment. This culminated in November 2005 when the implementation of the EPR was halted. The nurses and physicians of the surgical clinic called this situation a crisis. It was so major that some of the users hoped for discontinuation of the whole EPR implementation. The interviews in 2006 aimed to find out what kind of reasons and meanings were attached to the new technology and why its translation was so unsuccessful in the surgical clinic.

The emerging issues that caused rejection and resistance instead of adoption were related to problems related to the technology itself, to information needs and decision making in the clinic, and to organizational practices and culture. In all, 104 separate issues or complaints of the EPR were found during the analysis of the interviews (Forsell, Karsten & Vuokko 2007). Not all the issues were related to the EPR but there were also other socio-technical issues emerging from the organizational context of use. We could also note a practice of labelling the EPR as a “scapegoat” for various practical problems such as issues in re-defining organizational roles of responsibility. For example, there were many negotiations on who should document what and in what way. Recording responsibilities were closely intertwined with user accounts and passwords. The constant need to log on and off the EPR system was found rather frustrating when re-logging would always mean an interval of waiting the system to open up.

3.3 Maternity unit

The use of information technology in the maternity unit was explored during spring 2006. The maternity clinic, the birthing rooms and three maternity wards have 72 midwives working in three shifts. Each day, 15-20 babies are born. This part of data gathering involved observations in the unit over 4 months, following and tracking of 10 midwives, as well as interviewing 15 midwifes.

The study of the maternity unit took place when the midwives were still using Mama, the old character-based system for recording pregnancy, delivery and post-natal care, dating from the 1980s. They anticipated the new software i-pana to be implemented during autumn 2007. Their main problem was the assemblage of paper forms, entries into Mama and Miranda, some even to both, and separate programs for, for example, ordering laboratory tests. Especially the partogram that provided a quick overview onto how the delivery was progressing was seen as indispensable by midwives.

The mothers coming for delivery are closely monitored with various technologies, the well-being of the baby as the main focus. Monitoring the foetus's heartbeat together with the womb starting to open with contractions is vital for the timing of the actions by the midwives. In this context the organizational roles weighted in the translations of technology and delivery situations as especially the physicians and midwives have different expertise and adopt slightly different attitudes to birthing. It is noteworthy that in a Finnish hospital delivery, midwives have considerable authority and they are largely in responsible for women giving birth, while giving birth by section is not a common practice. The midwives were observed to care and protect the family in the delivery room, by even controlling access. Outside daily rounds, the obstetricians were largely absent, occupied in the clinic and invited only to perform some operations and attend to emergency cases.

3.4 The paediatric intensive care unit (ICU)

The work practices of ICU nurses were observed for 10 hours during the spring 2007 complemented with collecting ICU patient documentation sheets. The nurses also prepared us fictional control sheets and journal entries.

The paediatric intensive care unit (later ICU) is a small component in the overall complexity of a hospital, but in itself, it is a complex system that involves paediatricians, surgeons, and assisting physicians, anaesthesiologists, ICU-trained nurses, supporting staff, and multiple mechanical or electronic technologies. To control and simplify the collaboration between various participants the use of paper documentation of care has been a routine practice until the end of the year 2007 when the implementation of the EPR system was began in the unit.

In the ICU, for example, vitality monitors and other technical support systems are regularly used by the nurses. Patients' daily monitoring data is documented on paper forms that contain rich information and thus enable an overview of the patient's current situation.

An intensive care unit differs from an ordinary bed ward in that there is, in principle, one nurse per patient. However, the nurses collaborate much, especially in problem solving. The practice of "getting along within the unit" was so strong that we observed the ICU nurses applying the practice also on the adoption of formerly unfamiliar technological features.

4 Nurses translating technology

Nurses are one of the main user groups of the EPR system as they document, for example, patients' treatments, changes in medication, various test results and vitality monitoring data. New practices of documenting, collecting and sharing patient information in the EPR were negotiated and developed during the implementation, but there were also issues that emerged quite unpredicted and would challenge the objectives of the organizational implementation. One emergent issue was a co-construction of coordination as the nurses felt that they lose touch of their own working. In the new situation, a nurse could, for example, fill in a covering letter in the electronic form but didn't necessarily check that the intended receiver actually received it. A surgical clinic nurse evaluated in 2006 that problematic situations emerge when the work practices are rigid or inflexible:

> *"I think that these problems we have in the clinic, these are the real issues. In other units people have coped with the situation much better... for example in the internal diseases bed ward, there they are coping well... they don't have such rigid roles in their ward; they don't have roles inscribed in stone. They share the responsibilities and documenting, the person who has time or is capable makes the recording..."*

From the beginning of the organizational implementation one of the emerging issues has been positioning of patient information in the hospital wards and units. When thinking back, there has been recurrent issues concerning the placement and finding of patient information when the information was mostly gathered in patient paper folders. At that time, a patient folder could be found by the patient bedside, in a nurses' office or by a physician for reading, in the ward secretary's office waiting to be processed, or just on the way to the ward. Nurses assumed that with the EPR system, the issues of assessing patient information would be solved. In future, the patient information would be available on wall-mounted screens, on portable computers that could be moved around on a trolley, or on notepad computers. The patient information could be made assessable across the whole hospital, and would support local mobility (Bellotti & Bly, 1996) of the organizational actors in a working environment where the actors are constantly on the move to get ahead of their work tasks. Still, in the EPR, to see a patient's situation with one look on a screen is hard, and the users claim that using electronic records is slower than using a paper record. One reason for this is the structured character of information in the EPR. In the beginning of the implementation, in year 2003, a ward nurse assumed that structure on patient information would save time:

> *"The patient records will be well organized. There will be different parts and structures so that we'll easily find information even when the amount of data increases. With good headings we'll find the needed information quickly, and the information can be administrated efficiently."*

Usability problems and over 50 different headings structuring the data have caused the attitudes to change. In year 2006, a nurse from the surgical clinic described the hardships of use of the EPR in the following way:

> *"Now I have to open Miranda [EPR], to open the nursing records. Now I'll make the record, that takes many clicks – like surgeons name, date, and cause this and cause that. Then I'll have to choose the right headings, and then I can go and record the day visit by the patient, I can make it, and then I'll have to choose the next suitable heading… I have many phases here, phases that I have never done before… Before I just wrote, for example, 'covering letter' and 'breast cancer' n the paper and that was it."*

The slowness of using the EPR system affects the working in various ways. For example, in the surgical clinic there is continues patient visits during a working day, and half of the working time consists of making patient records. In contrast, documenting care in a bed ward takes only about 1/8 of the working time. In this light, it is understandable that the slowness of use can cause problems for daily work. This can cause the staff to think twice whether to use paper records of electronic records as they weigh the pros and cons of the EPR use – especially in situations where a patient has an acute need. For example, in the surgical clinic it was the nurses who decided that they will not use the EPR, whereas in the ward it was the physicians, who didn't want to use the EPR. This caused a situation where some of the patient records were on paper and some in the electronic form. In all, the work practices have been affected by the slowness of using the EPR right from the stars, as already in 2003 one of the nurses described the phenomena:

> *"Time is often tight, I feel that I could work faster but the program has a hard time keeping up, it stops to think often. Of course during the day there are peak times when many people are in the same part of the program... and are doing orders that must slow it down. Laboratory programs are always like that, many people use them around noon and often it is rather slow."*

One recurrent issue from the nurses' translations of the EPR was how the record keeping was reasoned. The nurses felt that the EPR doesn't so much help them in care work but that the system is in use to help the work of some third party, such as, hospital administration, health care researchers or makers of national health care statistics. In 2006, one of the nurses from the surgical clinic stated the interconnectedness between the EPR and practical nursing as follows:

> *"…it can be a little easier to find new information in the EPR, where they are structurally arranged than if we still would have the paper story… But in practical nursing it is often hard to understand why we do it in such a difficult way just because someone else wants to make statistics or research."*

More unified documenting practices have been planned within the organization, but still, system breakdowns due to technical problems have caused a situation where the documenting differs from what has been planned. A major source of uncertainty amongst the users of the EPR was the insecure technical infrastructure in the hospital. Especially the physicians are asking what happens if and when a patient's electronic records are not available. During breakdowns of the EPR use due to technical problems, there have emerged varied practices for handling the patient documentation.

For example, during breakdowns, patient records can be written as separate text files that are then later added in the EPR. Problems emerge when the separate text files are attached only as printouts to the paper version of the patient records and not in the EPR. This causes that the ER is not necessarily up-to-date, and that the staff members cannot trust the information in the EPR. Therefore, for example, the personnel in the surgical clinic felt that the use of the EPR can cause malpractice of the patients if care decisions are based on wrong or missing information content. To sum up, the expected mandatory use of the EPR and the breakdowns causing information gaps cause tension amongst the hospital staff.

An emerging translation of the new information technology suggests that instead of receiving more time for the care of the patients, the patient care practices are slowly changing to contain more and more interaction with the computers. In 2006, couple of nurses commented the recent developments as follows:

> *"Well, it's hard; we should speak with a patient to find out various things, but we have to concentrate our attention to fiddling with the computer... it is like, we talk to the computer, and the patient is forgotten."*
>
> *"Before, we nurses used to interact more with each others. We could, for example, process a care plan together... or write down patient documentation together. It was more interacting with each others, in a way, but now we just talk with the machine. In a way, we sit whole day by the computer and work with it."*

In contrast to bed ward nurses, the ICU nurses seemed to be familiar with different technologies supporting their care tasks. For example, when the ICU nurses prepare to take a new patient into the unit, they start the preparations by collecting needed equipments and vitality monitoring devices by the bed that is reserved to the new patient. The nurses know how to secure and attach the devices to right sensors and how to initialize the monitors according to the age of the patient. The preliminary alarms can be chosen from three levels that are an infant or a neonatal, a child or an adult. Besides patient monitoring the nurses prepare and monitor the IV fluids. Other preparations are possible, if the transfer information sheet implies such a need. Adjusting various machines is part of the routine, and the nurses are at most concerned with the right calculations but not on the functions of these machines. When more understanding about the inner function is needed, the nurses call for the support staff in charge of the technology. There are also expectations, as the physician is, for example, responsible for adjusting the breathing machine or the mechanical ventilator.

Especially the vitality monitoring technology is so embedded in the routines that the nurses claim to identify not only which device gives the alarm but also why it alarms. In reality, as one device has one alarm sound, a nurse interprets from other situational features why the alarm goes off. For example, she or he might have been already waiting for the alarm to sound so that the IV bag can be changed to a new one. On the other hand, when a patient is especially restless, such as, for example, a typical meningitis case, a nurse knows that the monitors set off alarms more than needed. If the nurse becomes uncertain in her or his interpretations or knowledge about the vitality monitors, a colleague is readily consulted and most solutions are formed together in the unit.

To sum up, as the nursing work practices are not only about sharing information but also about sharing responsibility of the patients, the nurses are also jointly constructing the interpretations attached to the new technology as they try attaching the EPR with meanings that would make it a part of nursing work. While the nursing documenting practices are the only visible trace of the care, the efforts conserved to the process of translation can be understood as the EPR is accounted to have a considerable effect on care work.

5 Conclusions

Within the process of translating technology, the nurses formed a number of different interpretations for the use of the EPR. The varied interpretations were emerging from the different realities or emphasis of care across the many hospital units, and the variation affected the adoption of the EPR. For example, a bed ward nurse and an ICU nurse have different tasks or objectives in care practices. While the nurses have different needs and rapidity for patient records, the EPR system is used to support varied tasks within the same hospital.

In the light of our four study cases, the organizational implementation seemed to raise a number of issues that were related to the hospital organization and to the coordination of care work. These issues included, for example, positioning of information technology in the wards and clinics, the division of tasks or role descriptions and organizational responsibilities, and interdependences of documenting practices.

The structured patient record was interpreted as slow to use and as "hiding" the patients from nurses' gaze. The hospital nurses recalled that the paper sheets for recording patient information supported nursing practices well in the sense, that a paper sheet could be quickly glanced, and the detailed patient information on a sheet was constantly updated by small notes. The view on the care situation would emerge bit by bit on the paper sheet in a way that reminded the nurses of an oil painting where a layer upon layer of colour and details can be added.

Usability problems unfolding as slowness of use in daily recording of care and breakdowns with related technical problems caused further uncertainty of interpretation amongst the nurses. Uncertainty or unsuccessful translation of technology emerged as user resistance that peaked in the halting of the organizational implementation in the surgical clinic. Disappointment to the EPR in use was further caused by the interpretations that the implemented technology was something extra, something mandatory that didn't support nurses' daily working in a sufficient manner.

The hospital case indicates that solely technical issues, such as usability problems, can cause user resistance or uncertainty. On the other hand, issues concerning professional values, such as fear of malpractice due to missing or un-accessible patient information, can cause a decision not to use the system at all. Noteworthy is that the reason for resistance was not purely technical or social issue but a combination of socio-technical issues that emerged during the organizational implementation.

Transforming work practices and interpreting new uses from technology often take place when, for example, the current work practices no longer respond to the technical tools in use. Through the adaptation to changing situation, not only new technology is being taken into use but at the same time, new work practices as well as new contextual knowledge may emerge.

6 Acknowledgements

The interviews and observations in the four locations described here were carried out by our Fulbright visitor for Spring 2006 professor Kent Sandoe; our Master's students Aija Laine, Sami Laine, Mika Koivunen, Katariina Sovela, Anne Forsell, Joonas Peltola, Pekka Tetri; the students of the Current Issues in Information Systems course in 2001-2; and ourselves. Funding for our work was provided by TEKES grants 40679/01, 40435/05 and 40020/07.

References

Bellotti, V., and Bly, S. (1996). Walking away from the Desktop Computer: Distributed Collaboration and Mobility in a Product Design Team. In Ackerman, M.S. (Ed.) Proceedings of the Conference on Computer Supported Cooperative Work 1996, 209-218.

Berg, M. (2004). Health Information Management: Integrating information technology in health care work. Routledge, London.

Berg, M. (2001). Implementing information systems in health care organization: Myths and challenges. International Journal of Medical Informatics 64, 143-153.

Bijker, W.E. (1995). Of Bicycles, Bakelites, and Bulbs. Toward a Theory of Sociotechnical Change. The MIT Press, Cambridge, MA.

Ciborra, C.U. (2002). The Labyrinths of Information: Challenging the Wisdom of Systems. Oxford University Press, Oxford.

Cohen, M. (1999). Commentary on the Organization Science Special Issue on Complexity. Organization Science 10(3), 373-376.

Doolin, B., and Lowe, A. (2002). To Reveal is to Critique: Actor-Network Theory and Critical Information Systems Research. Journal of Information Technology 17, 69-78.

Forsell, A., Karsten, H. and Vuokko, R. (2007). Organizational implementation in crisis: The role of the information system, IRIS Conference, 11.-14.8.2007, Tampere, Finland.

Gallivan, M. (2001). Organizational Adoption and Assimilation of Complex Technological Innovations: Development and Application of New Framework. The DATABASE for Advances in Information Systems 32(3), 51-85.

Haythornthwaite, C. (2006). Articulating Divides in Distributed Knowledge Practice. Information, Communication & Society 9(6), 761–780.

Jacucci, E., Hanseth, O., and Lyytinen, K. (2006). Introduction: Taking Complexity Seriously in IS Research. Information, Technology & People 19(1), 5-11.

Jeyaraj, A., Rottman, R., and Lacity, M. (2006). A Review of the Predictors, Linkages and Biases in IT Innovation Adoption Research. Journal of Information Technology 21(1), 1-23.

Jones, M. (2003). Computers can land people on Mars, why can't they get them to work in a hospital? Implementation of an Electronic Patient Record System in a UK hospital. Methods of Information in Medicine 42(4), 410-415.

Karsten, H., and Laine, A. (2007). User Interpretations of Future Information System Use: A Snapshot with Technological Frames. International Journal of Medical Informatics, Volume 76, Supplement 1.

Knox, H., Savage, M., and Harvey, P. (2006). Social Networks and the Study of Relations: Networks as Method, Metaphor and Form. Economy and Society 35(1), 113-140.

Latour, B. (2005). Reassembling the Social. An Introduction to Actor-Network-Theory. Oxford University Press, Oxford.

Latour, B. (1999). Pandora's Hope. Essays on the Reality of Science Studies. Harvard University Press, Cambridge, MA.

Latour, B. (1991). Technology is Society Made Durable. In Law, J. (Ed.), A Sociology of Monsters: Essays on Power, Technology and Domination (pp. 103-131). Routledge, London.

Leidner, D.E., and Kayworth, T. (2006). A Review of Culture in Information Systems Research: Toward A Theory of Information Technology Culture Conflict. MIS Quarterly 30(2), 357-399.

Merali, Y. (2004). Complexity and Information Systems. In Mingers, J., and Willcocks, L. (Eds.), Social Theory and Philosophy for Information Systems (pp. 407-446). John Wiley & sons, Ltd, Chichester.

Middleton, D., and Brown, S.D. (2005) Net-working on a Neonatal Intensive Care Unit: The Baby as Virtual Object. In Czarniawska, B., and Hernes, T. (Eds.), Actor-network Theory and Organizing (pp. 307-328). Liber & Copenhagen Business School Press, Malmö.

Moser, I. (2005). Information and its uses in medical practice: a critical interrogation of IT plans and visions in health care. International Journal of Action Research 1(3), 339-373.

Moser, I., and Law, J. (2006). Fluids or flows? Information and Qualculation in Medical Practice. Information, Technology & People 19(1), 55-73.

Orlikowski, W.J. (2002). Knowing in Practice: Enacting a Collective Capability in Distributed Organization. Organization Science 13(3), 249-273.

Rogers, E. (2003). Diffusion of Innovations (5th edition). The Free Press, New York.

Scott, J.T., Rundall, T.G., Vogt, T.M., and Hsu, J. (2005). Kaiser Permanente's experience of implementing an electronic medical record: a qualitative study. British Medical Journal 331, 1313-1316.

Star, S.L., and Ruhleder, K. (1996). Steps towards an ecology of infrastructure: design and access for large scale information spaces. Information Systems Research 7(1), 111-134.

Taylor, S.J., and Bogdan, R. (1998). Introduction to Qualitative Research Methods, 3rd Edition. Wiley, New York.

Vuokko, R. and H. Karsten (2008). Transforming work practices in a complex environment. Information Technology in the Service Economy: Challenges and Possibilities for the 21st Century. M. Barrett, E. Davidson, C. Middleton and J. I. deGross (Eds.). Boston, Springer: 143-157.

Wainwright, D.W., and Waring, T.S. (2007). The application and adaptation of a diffusion of innovation framework for information systems research in NHS general medical practice. Journal of Information Technology 22(1), 44-58.

Woolgar, S. (1991). Configuring the user: the case of usability trials. In Law, J. (Ed.), A Sociology of Monsters: Essays on Power, Technology and Domination (pp. 57-101). Routledge, London.

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Short Papers

Short papers are double-blind peer-reviewed and they are presented as a poster.

The Use of Structured Documentation in Electronic Patient Records: Case Diabetics

Virpi Jylhä[1], Ulla-Mari Kinnunen[2] and Kaija Saranto[1]
[1] University of Kuopio, Department of Health Policy and Management, P.O. Box 1627,
70211 Kuopio, Finland
[2] Kuopio University Hospital, PL 1777, 70211 Kuopio, Finland
virpi.jylha@uku.fi, ulla-mari.kinnunen@kuh.fi, kaija.saranto@uku.fi

Abstract. The use of electronic patient records (EPRs) in Finland has become more widespread during recent years and thus had an effect on the way nursing care has been documented. The structure of nursing documentation is based on the nursing process model and nursing diagnosis, interventions and outcomes are documented using a standardised nursing terminology. Patient related information is produced and stored in electronic form at multiple sites. These databases enable evaluation, analysis and utilisations of data for administrative and research purposes. The purpose of this study was to describe analysis and utilisation of electronically extracted data from electronic patient records. In this study the analysed data were taken from the EPRs and nursing information was combined with diagnosis information of diabetics. The medication-related data of diabetics were analysed by using quantitative and qualitative methods. According to the results nursing diagnoses were mainly documented with narrative text whereas nursing interventions were documented by using medication component. As a conclusion EPR data is an excellent source of information in improving patient care and more information is needed for how technology can be used to create innovative approaches to integrating clinical data into practise.

1 Introduction

The use of electronic patient records (EPRs) has become more widespread during recent years. [1] This has had an effect on the way patient care is documented as well as what kind of care-related data is available for decision-making and research.

In tandem with the development of electronic health record systems technological progress in data acquisition and storage capacities has resulted in huge databases in health care. [2] Patient related information is produced and stored in electronic form at multiple sites. These databases enable evaluation, analysis and utilisations of data for administrative and research purposes. [3]

The purpose of this study was to describe analysis and utilisation of electronically extracted data from electronic patient records. Further, the aim was to find out what kind of information electronic databases offer for research purposes when structured nursing documentation is used. A case-study is presented to illustrate the analysis of medication documentation of diabetes patients.

2 Electronic Structured Documentation in Nursing

Clinical patient data can answer a variety of questions presented by managers, researchers or policy makers when it is collected and used appropriately. Recently, research using electronically stored clinical data has become increasingly common and effective. Documentation developments such as the increased standardisation of patient record forms and use of classifications have made healthcare data a more reliable and useful resource for health research. [3]

When data mining methods are used to convert nursing data to knowledge, standardised documentation is a prerequisite. In Finland, electronic structured documentation is presented nationwide and has changed the ways in which information is produced and utilised. Nursing documentation content can be structured by means of the nursing process model, which consists of five phases: assessment, diagnosis, planning, implementation and evaluation. Classifications and terminologies should be used in order to standardise the structure of documentation. [8] When nursing diagnoses, interventions and outcomes are documented consistently, the documentation will produce descriptive information about patients' problems and about nursing interventions, allowing decision making at clinical, administrative and policy levels. [4], [5]

In Finland, the classifications have been implemented in the EPRs and are used for describing nursing diagnoses and interventions. The Finnish Care Classification (FinCC) is based on the Clinical Care Classification (CCC) developed in the USA. [6] The Finnish Classification of Nursing Diagnosis (FiCND) and Interventions (FiCNI) are both included in FinCC. In the data used in this paper, FiCND version 1.0 and FiCNI version 1.2 were used to document patient care. Those versions contain 17 main components; these are divided into a number of main categories and further into subcategories. [7], [8]

3 Materials and Methods

In this paper the focus is the medication information of diabetics extracted from the EPR system. Under the Finnish law, the need for medication care, prescriptions and the administration of medication have to be documented in the patient records. [9] The Finnish Classification of Nursing Diagnosis is used to describe actual and potential health problems of patients and the Finnish Classification of Nursing Interventions Medication component is used to describe nursing interventions. When documenting e.g. the need for medication, the medication component with its main and subcategories are used. The nursing diagnosis and interventions related to medication can be documented with or without narrative text. [7], [8]

In this study the analysed data were retrieved from the EPRs of one central hospital in Finland in 2005. The permission to carry out the research was admitted by the hospital research council. The data used in this study included structured documented patient data from the medical wards. One database in SQL-format was constructed from EPR data. Each data source, i.e. nursing diagnosis, nursing intervention and medical diagnosis, were transmitted to tables included in the research database. In the nursing diagnosis and nursing intervention tables each patient record included the patient’s number, the title of the user, the heading, the date and time, narrative text and the FiCNI or FiCND codes. The diagnosis table consisted of demographic data i.e. patient's age and year of birth, place of living and gender as well as the diagnosis codes of the patient.

The patient’s identification numbers were transformed into an anonymous form in the hospital before the data were delivered to the researchers on a CD-ROM. The data were stored in SQL format in the research database and MS Access was used to combine tables and classify the data. The data were analysed by using quantitative methods. MS Access and MS Excel programs were used in the descriptive analysis.

Nursing information was combined with diagnosis information. The use of classifications were analysed as well as the documentation of narrative text. The search of nursing data for patients who have diabetes mellitus was limited to ICD 10 – codes E10-E14. In addition to gather documentation concerning nursing diagnoses and interventions, the searches were conducted in the narrative text by using the word "medication" and its different forms in the Finnish language.

4 Results

4.1 Nursing Diagnosis

The Finnish Classification of Nursing Diagnosis to describe patients’ needs was used totally in 1669 records on medical wards in 2005. Medication component was used totally 105 times. Further, only 19 records of patients with diabetes were included in one of the categories of the medication component. The most used main category for diabetics was Knowledge Deficit of Medication Regimen, which is also the most used category on medical wards. (Table 1.)

A qualitative analysis of narrative text of diabetics showed that in 8 records (n=19) only the title was used, without specification of nursing diagnoses. In addition, the use of main category did not correspond with narrative text in 7 records (n=19). For example, patient's status, implemented interventions or list of medication instead of patient's needs for information were described in the narrative text, though the main category Knowledge Deficit of Medication Regimen was used.

Nursing diagnosis are mainly (2148 times out of 2246) documented with narrative text and without any main or sub categories of medication component. The result is the same for diabetics in which case most of the records (313 out of 330) were documented by using narrative text without component. (Table 2.)

4.2 Nursing Interventions

The records of the medical wards included in total 21 891 records in which the Medication component was used at least once in 6056 records and totally 6683 times. For diabetics the medication component was used a total of 994 times, and at least once in 905 items. (Table 3.)

Nursing interventions are mainly documented by using narrative text with medication component. In the narrative text the heading "medication" was used for 6713 records. Out of these, 704 records were documented without use of the Medication component. For diabetics the heading "medication" was used for 993 records, and 97 records were documented without use of classification. (Table 4.)

5 Discussion

Nurses believe that their clinical practice is in contact with nursing records [10], but the reality is not always so. As our case-study has shown narrative documentation and used classification are described inconsistently. This indicates deficiencies in knowledge about classifications but also inaccurate documentation. The study made by Bakken et. al (2007) confirms our results. According to their results the quality of the information concerning medication was unreliable and documented diagnoses were not accurate enough. [11]

To obtain reliable results, in addition to the quality of the data, the methodology of data extraction needs to be considered carefully. Since patient data were not originally designed for research, there are obvious limitations to using these data [12]. Consequently, there are numerous steps taken to transform and clean these data for research purposes. Nevertheless, since these data are often the only source available, data manipulation techniques have been developed to optimise the use of this data. This increases the probability of obtaining a data set with a higher degree of reliability and validity. [13] To guarantee reliable results, the understanding of health care and the content of clinical documents is required. In addition, it is essential to understand the structure of EPRs when extracting the data for research purposes.

Data mining brings new tools and possibilities for analysing data in health care. By utilising electronic warehouses, the activities in health care and possible research subjects can be brought forth almost without delay. That is if the data in warehouses can be processed to knowledge. Also, accretion of data and knowledge from nursing practices enables the development of, from the patient’s point of view, better quality health care services. [14, 15]

Evaluation of documentation and use of nursing information is much easier when using terminologies in electronic patient record systems. Firstly, the necessary data are easy to gather directly from the electronic databases and secondly, electronic patient record systems allow for the inclusion of a wider range of data items in the analysis. When information extracted from electronic databases is used to evaluate nursing practise or decision-making, the basic assumption is that the data is accurate. The results of the case-part of this study show that the use of classifications is partly unreliable. Also Muller-Staub et all. (2006) indicated that the accuracy of nursing documentation needs improvement [16].

When using data from the electronic patient records for other purposes than patient care, privacy and human rights concerns as well as legality and ethics needs to be considered [15]. In health care, data mining provides information that is difficult to obtain otherwise, as this data, compiled from electronic patient records, is an important information resource for analysis of current health care practices.

In conclusion, EPR data is an excellent source of information in improving patient care and more information is needed for how technology can be used to create innovative approaches to integrating clinical data into practise. The issues concerning the quality of electronically stored clinical data should also been considered, before using the data for other than clinical purposes.

References

1. Hämäläinen, P., Reponen, J., Winblad, I.: eHealth of Finland: Check point 2006. Stakes Reports 1/2007, Valopaino Oy (2007)
2. Hand, D., Mannila, H., Padhraic, S.: Principles of Data Mining. The MIT Press, Cambridge (2001)
3. Perry, T.L., Tucker, T., Hudson, L.R., Gandy, W., Neftzger, A.L., Hamar, G.B.: The Application of Data Mining Techniques in Health Plan Population Management: Disease Management Approach. In Kudyba, S. (ed.). IT Solutions Series: Managing Data Mining: Advice from Experts. pp. 135--153. Hershey, PA, USA, Idea Group Inc (2004).
4. Ensio, A. Saranto, K.: Hoitotyön elektroninen kirjaaminen. [Electronic Documentation of Nursing] Silverprint, Sipoo (2004) In Finnish.
5. Goossen, W.T.F., Epping, P.J.M.M., Dassen, T.: Criteria for Nursing Information Systems as a Component of the Electronic Health Record. Comput Nurs 15(6), 307--315 (1997)
6. Saba VK. Clinical Care Classification System, http://www.sabacare.com.
7. Hoitotyön toimintoluokitus [The Finnish Classification of Nursing Interventions], http://194.89.160.67/codeserverTES/classification-action.do?action=find&key=89. In Finnish.
8. Hoitotyön tarveluokitus [The Finnish Classification of Nursing Diagnoses], http://194.89.160.67/codeserverTES/version-action.do?action=find&version=235. In Finnish.
9. Law 785/1992. Act on the Status and Rights of Patients. (in Finnish).
10. Currell R. & Urquhart C. Nursing record systems: effects on nursing practice and health care outcomes. Cochrane Database of Systematic Reviews, Issue 3. Art. No.: CD002099. DOI: 10.1002/14651858.CD002099, (2003)
11. Bakken, K., Larsen, E., Lindberg, P. C., Rygh, E., Hjortdahl, P.: Mangelfull kommunikasjon om legemiddelbruk i primærhelsetjenesten [Insufficient communication and information regarding patient medication in the primary healthcare sector] Tidsskr Nor Lægeforen 127, 1766--1769 (2007)
12. Magee, T., Lee, S.M., Giuliano, K.K., Munro B.: Generating new knowledge from existing data. The use of large data sets for nursing researh. Nursing Research 55(2S), 50--56 (2006)
13. Larose, D. T.: Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Incorporated, Hoboken, NJ, USA (2005)
14. Shever, L.L., Titler, M., Dochterman, J., Fei, Q., Picone, D.M.: Patterns of Nursing Interventions Use Across 6 Days of Acute Care Hospitalization for Three Older Patient Populations. Int J Nurs Terminol Classif 18(1), 18--29 (2007)
15. Burns, N., Grove, SK.: The Practise of Nursing Research. Conduct, Critique & Utilization. 5th ed. W.B. Saunders Company, the United States of America (2005)
16. Müller-Staub, M., Lavin, M.A., Needham, I., van Achterberg T.: Nursing diagnoses,interventions and outcomes – application and impact on nursing practice: systematic review. Journal of Advanced Nursing 56(5), 514--531 (2006)

Appendix: Tables

Table 1. Nursing Diagnosis: Use of the main and subcategories of the Medication component

MAIN CATEGORY	All patients n=106 records	Patients with diabetes n=19 records
G.1 Medication risk	19	1
G.1.1 Polypharmacy	4	0
G.1.2 Unsuitability of Medication Regimen	6	0
G.2 Knowledge Deficit of Medication Regimen	76	18
Total	**105**	**19**

Table 2. Use of narrative text to describe nursing diagnosis

NARRATIVE TEXT	All patients	Patients with diabetes
In addition to use of Medication component G	98	17
Without use of Medication component G	2148	313
Total	**2246**	**330**

Table 3. Nursing Interventions: Use of the main categories of the Medication component

MAIN CATEGORY	All patients* n=6056 items	Patients with Diabetes n=905 items
G.1. Medication Administration	6191	923
G.2. Medication Side Effects	84	4
G.3. Medication Counselling	408	67
Total	**6683**	**994**

Table 4. Use of narrative text to describe nursing interventions

NARRATIVE TEXT	All patients	Diabetic patients
In addition to use of Medication component G	6009	896
Without use of Medication component G	704	97
Total	**6713**	**993**

From Clinical Notes to ICD: Development of a Coding Help

Julia Medori

CENTAL
Université catholique de Louvain
Place Blaise Pascal, 1
1348 Louvain-la-neuve
julia.medori@uclouvain.be

Abstract. This paper aims at describing an automatic encoding system which reads clinical notes and encodes the information into the Clinical Modification of the International Classification of Diseases, 9th revision (ICD-9-CM). This work is done in collaboration with a university hospital in Brussels. The system is structured into two steps: extraction and codification. The extraction process has been fully implemented and will be integrated into the hospital's current coding software as a coding help.

1 Introduction

Mapping text into the International Classification of Diseases (ICD) or other classifications for the purpose of indexing documents or helping coders has been the subject of numerous studies: [2], [3], [4], [5], [7], [8], [9], [12], [13]. Our work may be distinguished from these studies: first it is in French and therefore benefits from fewer resources than in English; our system is designed to work on clinical notes from all medical services whereas previous studies often chose to focus on one of them; and our source files are not structured documents (no structure tags or titles). Automatically encoding clinical notes into ICD is of great interest to hospitals and physicians as it could greatly reduce the time spent on paper work. However, these systems have long been studied but few have been fully implemented.

This paper aims at presenting an empirical implementation of an encoding system. This project is the product of a collaboration between the “Cliniques universitaires St-Luc”, a hospital in Brussels and the Centre for Natural Language Processing (CENTAL) at the University of Louvain (Belgium). Currently a team of file clerks is in charge of manually encoding clinical notes into the ICD-9-CM, which is the classification currently in use in this hospital. This encoding process is vital to hospitals as the codes are analyzed by the Federal Public Service in order to evaluate hospital’s fundings. This work is tedious and time-consuming and as administrative work is becoming more and more demanding, there is a growing need for a tool that will help quicken and ease this task.

2 Automatic Encoding

Previous works have proved that this goal can be achieved. A few systems mapping text and codes have been implemented as coding helps or document indexing tools [8]. Some of them are in use in hospitals where they became essential tools. In [1], the system architecture of Biomedical text mining tools is pointed as fairly standard and generally comprises three steps:

- Lexical analysis
- Syntactic analysis
- Semantic analysis

This general pattern has to be considered in light of the task at hand. The main characteristics of our source files, the clinical notes, is that they are all written in free text. Moreover, these letters are often written in a short amount of time and seldom contain full sentences. The style is telegraphic and does not strictly comply with French syntax and this makes us question the reliability of a syntactic analysis.

The Medical NLP Challenge [10] organized by the Computational Medicine Center at the start of the year 2007 highlighted the different approaches chosen by different teams working on this issue. The goal of this challenge was to assign ICD-9-CM codes to radiology reports written in free text. However the corpus had been preprocessed in order to allow a statistical approach: the main articulations of the text were marked up and the range of possible codes considerably reduced to 45 items. Each combination of codes appeared at least twice in the corpus.

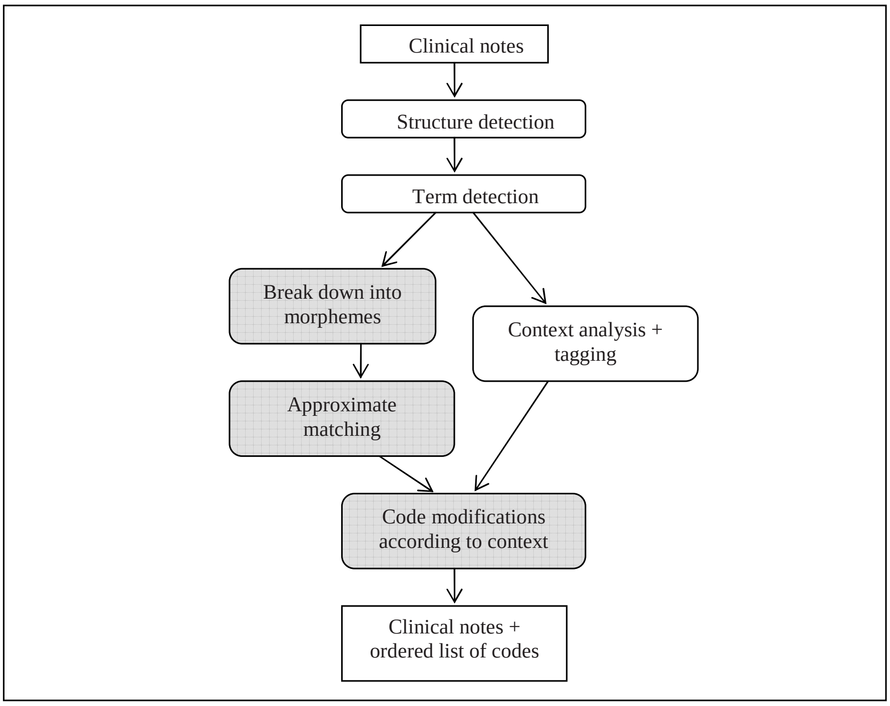
Among the best three systems, two combined a statistic and a symbolic approach and only one relies only on a symbolic approach. Most systems participating took a hybrid approach.

During ACL 2007, Aronson [2], presented within the framework of the same challenge, four different approaches, symbolic, statistical and hybrid. His conclusion were that combining different methods and approaches performed better and were more stable than their contributing methods.

3 System Architecture

In our case, we chose to focus first on a symbolic approach to deal with this issue. We will consider combining it with a statistical approach at a later stage. The architecture of the system is shown in figure 1. Our method comprised two steps: at first, important terms are extracted and the context in which they occur is analyzed, then these terms are compared to the codes of the classification. These steps are described in this section.

Fig. 1. System structure


Clinical notes
Structure detection
Term detection
Break down into morphemes
Context analysis + tagging
Approximate matching
Code modifications according to context
Clinical notes + ordered list of codes

3.1 Extraction

The aim of the first step: the extraction process, was to read clinical notes and extract the terms containing the information that is necessary for the encoding process. These terms generally refer to diseases but also include anatomical terms, the degree of seriousness or probability of a disease, or other types of information that may influence the choice of a code.

As pointed out in section 1, the main characteristics of our source files are:

- **Free text**: they are written in the form of a letter addressed to the patient's GP. There are very few predefined fields to be filled in.
- **Variation in structure**: The letter is often divided into sections. However this partition does not rely on any accepted structure. Therefore the structure of the letters vary according to medical services and writers. A close look at our corpus enabled us to derive a general structure: introduction, past history, present history, exams, conclusion.
- **Telegraphic style**: a very short amount of time is dedicated to writing these notes. Therefore, sentences are often limited to very short phrases that do not strictly comply with standard French syntax.
- **Variation in length**: According to the service or physician, the amount of information contained in a letter varies. A letter from the hematology service will often be very long (~4 pages) and very detailed whereas a letter from plastic surgery will be extremely short (~10 lines).

These characteristics are important to the extraction process as they influence the structure of the system.

Detecting the way the letter fits into the general structure is the first step of the extraction. Knowing which terms belong to which section will be helpful to the encoding process: a disease appearing in the "past history" section may not be active at the time of writing and will not be encoded, however if it appeared in the conclusion section then it is most likely to be encoded. This detection process is based on a keyword search.

The second step marks up the terms that refer to diseases, body parts, degree of seriousness, probability of presence of the disease, drugs and other terms that may influence the encoding process. This process uses "home-made" dictionaries for locating these terms. The main difficulty one has to face when building such a system for French is the lack of biomedical resources compared to English. The first step was then to collect all the data we could in order to build our dictionaries of medical terms. For the purpose of building our dictionary of diseases, we mainly used the ICD-9-CM itself and the French classifications included in the UMLS.

Once these terms are extracted, detecting in which context they occur is essential: a disease will be considered differently when negated, probable or present. This was done thanks to finite-state automata and transducers that described grammars of the language used in clinical notes. According to their context, the extracted terms are marked up with XML tags as being present, absent, past history, probable etc. A set of synonyms and abbreviations of the dictionary terms is also used. It allows the system to recognize and link these synonyms to their original forms, the one used in the classification. The analysis of the context also detects clues in the text that point to a term that may not be in our dictionaries but which may be important to the encoding process. For instance, "Troubles de … " (equivalent to "… disorder" in English) indicates an abnormality or a disease that might need to be encoded.

A module designed to detect the patient's height and weight is also added to this extraction process. This allows the system to compute the body mass index. When non-standard, this piece of information must always be encoded.

3.2 Codification

The steps involved in the codification process are the greyed areas on figure 1. This process takes into account the terms extracted in the previous phase and compares them to the codes of the ICD-9-CM. This part of the work is still in progress but a general structure has been developed.

From a close look at the terms used by physicians compared to the ones used in the classification to refer to the same disease, we noticed that they often were morphologically close.

Fig. 2. Example of cervical arthrosis

Cervicarthrose (physician)	Ostéoarthose cervicale (ICD-9-CM)
Cervic-	Ostéo-
Arthrose	**Arthrose**
	cervic-
	-al(e)

A module inspired by the works of Namer [6] was developed. It breaks down the extracted terms into morphemes. It was therefore included as a first step in the codification process. Figure 2 presents the example of two ways of refering to cervical arthrosis. In this example, two morphemes are common to both wordings: cervic- and -arthrose. However, morphemes are not enough. "Dorsalgia" for example may be refered to as "back pain" which is the literal meaning of each morpheme: *dors-* and *-algia*. The list of morphemes was therefore extended with the meanings of each morpheme. The module then computes a similarity value between the extended term from the letter and the codes in the classification extended in the same way. The code with he highest score will therefore be assigned to the extracted term.

The last step filters out the codes according to the context detected during extraction.

4 Evaluation

A first evaluation of the extraction process has been carried out. Its goal was to evaluate the performance of the system at highlighting "important terms" compared to what human coders would consider as important. It consisted in asking human coders to manually mark up the information they translate into codes. 8 coders from the "Cliniques universitaires St Luc" took part in this evaluation. The corpus comprised 220 letters: 10 letters for each of 22 medical units. These letters were then analyzed by the system. The evaluation of extraction systems usually involves the measure of precision and recall. 3700 terms or phrases were manually extracted. 66.6% of these phrases were also extracted by the system (recall). However, it is difficult to measure precision in the same way: the automatically extracted terms that were not also manually extracted are not always errors. The main difference between the system and coders is that the system tries to be exhaustive in its search for diagnoses and symptoms where coders will only look for the main diagnoses and will not consider symptoms as relevant when the associated disease is clearly mentionned. 5673 elements were extracted by the system. Among these, 8.1% can be considered as errors. They were often due to the tool that extracts potential diseases missing in the dictionary, thanks to clues in the context (see section 3.1). This step is essential as it retrieves many diseases and symptoms but the noise will need to be filtered out. 6.4% of the automatically extracted terms were wrongly classified according to context. More works on the graphs would lower this value. However, some letters contain predefined fields such as "inflammatory syndrome". The physician is expected to fill in these fields with a cross, a plus or minus or other signs indicating the presence of the disease. Many physicians leave it blank when the diagnosis is absent. The system assumes that when the context does not show any sign of absence or likelihood, the disease is present. This leads to many errors and a filter should be built for these types of letters. It should be stressed that this is a first evaluation of the extraction system. These results are promising as with the development of filters and the improvement of graphs and dictionaries, we may greatly reduce the noise. In order to further improve our results, we will consider combining our approach with statistical methods.

An interface is in development in order to integrate this tool into the software currently used to fill in the coding form: Acodam [11]. At this stage of the development, we are planning on providing a coding help by allowing coders to see the clinical notes with the extracted terms highlighted. However, not all of them will be highlighted as not to overcome coders with irrelevant information. They will need to be filtered out to avoid highlighting multiple occurrences and according to the context in which they occur.

5 Conclusion

Being able to automatically encode clinical notes would be a significant achievement for hospitals. We detailed in this paper a linguistic approach of this problem. Our method was divided into two steps: extraction and codification. The extraction process has been fully implemented and will soon be integrated into the coders current software as a coding help.

Acknowledgements

The CAPADIS project was funded by the government of the Brussels-Capital Region (Institute for the encouragement of Scientific Research and Innovation of Brussels) within the framework of the "Prospective research for Brussels 2006" program. I would also like to thank the St Luc hospital team of coders who helped me carry out the evaluation.

References

1. Ananiadou S., McNaught J.: Introduction to Text Mining in Biology. In Ananiadou S., McNaught J. (eds.) Text Mining for Biology and Biomedicine, pp 1--12, Artech House Books.(2006)
2. Aronson A. R.: MetaMap: Mapping Text to the UMLS Metathesaurus.(2006)
3. Ceusters W., Michel C., Penson D., Mauclet E.: Semi-automated encoding of diagnoses and medical procedures combining ICD-9-CM with computational-linguistic tools. Ann Med Milit Belg;8(2):53--58.(1994)
4. Deville G., Herbigniaux E., Mousel P., Thienpont G., Wéry M.: ANTHEM: Advanced Natural Language Interface for Multilingual Text Generation in Healthcare (1996)
5. Friedman C., Shagina L., Lussier Y.A., Hripcsak G.: Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392--402. Epub 2004 Jun 7. (2004)
6. Namer F. : Morphosémantique pour l'appariement de termes dans le vocabulaire médical: approche multilingue. TALN 2005 (6-10 juin 2005), Dourdan, pp.63-72.(2005)
7. Névéol A., Rogozan A., Darmoni SJ. : Automatic indexing of online health resources for a French quality controlled gateway. Information Processing and Management, May 2006 - 42:3 :695--709. (2006)
8. Pakhomov S. V. S., Buntrock J. D., Chute C. G.: Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-**based** and Machine Learning Techniques. (2006)
9. Pereira S., Névéol A., Massari P., Joubert M., Darmoni S.J. : Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. Proc. MIE. 2006.
10. Pestian J. P., Brew C., Matykiewicz P.M., Hovermale D.J., Johnson N., Cohen K.B., Duch W.: A shared task involving multi-label classification of clinical free text. Proceedings of ACL BioNLP; 2007 Jun; Prague. (2007)
11. Roger France F. H., Beguin C., De Clercq E. : Systèmes d'information pour la typologie des malades en Belgique. In Journées Francophones d'Informatique Médicale (JFIM2005), 1-7, Lille, France. (2005)
12. Sager N., Lyman M., Nhán N., Tick L.: Medical language processing: Applications to patient data representation and automatic encoding. Methods of Information in Medicine, (34):140 -- 146. (1995)
13. Zweigenbaum P. and Consortium MENELAS: MENELAS: coding and information retrieval from natural language patient discharge summaries. In Laires M. F., Ladeira M. J., Christensen J. P., (eds.), Advances in Health Telematics, pages 82-89. IOS Press, Amsterdam, 1995. MENELAS Final Edited Progress Report. (1995)

Domain-specific Analytical Language Modeling – the Chief Complaint as a Case Study

Samuli Niiranen[1], Michael Aswell[2], Jari Yli-Hietanen[1],
and Larry Nathanson[3]

[1]Department of Signal Processing, Tampere University of Technology,
33101 Tampere, Finland
[2]AthenaHealth Inc., Watertown, MA 02472, USA
[3]Department of Emergency Medicine, Beth-Israel Deaconess Medical Center,
Harvard Medical School, Boston, MA 02115, USA

{samuli.niiranen, jari.yli-hietanen}@tut.fi, michael.aswell@athenahealth.com,
lnathans@bidmc.harvard.edu

Abstract. We discuss the need for domain-specific analytical modeling in the computational understanding of medical text. To exemplify this, we present a case study and evaluation results using a specific class of medical text, namely emergency department chief complaints.

1 Introduction

Natural language processing is an important tool in medical informatics. A large share of the information in electronic medical records (EMRs) consists of free-text compositions. From a computational point-of-view, the continuing prevalence of free-entry is a major hindrance when the goal is to increase automation in EMRs. However, the efforts in developing standards for the structured representation of medical information have not proven to be a panacea. The information space of clinical medicine is very diverse and constantly evolving making it challenging to develop standards for the domain.

The syntax and semantics of the natural language used in medical free text often differs radically from those used in standard, everyday language perhaps best exemplified by that used in newspaper articles [1]. The observed differences include:

- Use of highly specialized and obscure vocabularies
- Use of abbreviations with non-standard expansions
- Use of radically compressed and non-standard phrase and clause structures
- The prevalence of misspellings, concatenations, etc. in some medical texts

Furthermore, there is significant diversity in the ways in which language is used to convey meaning also among different medical specialties and providers. This often results in the need to develop narrowly domain-specific analytical tools for medical natural language processing, especially if purely statistical tools can't provide the required performance. Also, as noted earlier, due to the diverse and evolving information space, clinical NLP tools should be easily and efficiently trainable.

2 Case Study: Emergency Department Chief Complaints

2.1 Background, Motivation and Goal

The chief complaint (CC) is one of the most important components of emergency department (ED) triage decision making. It not only is essential in determining the direction of physical examination and diagnostic testing, but also serves as a historical document in the patient's clinical history [2]. Underlying its importance is the fact that the CC is the first available description of what's wrong with the patient. The CC, as reported by the patient in his or her own words, is typically recorded as a free-text entry to the ED EMR.

A number of studies (e.g., [3]) report activities towards a standard, limited set of encodings for the structured entry of chief complaint information in ED EMRs. However, no consensus exists on this matter and it remains a challenging and elusive goal due to open and evolving nature of the underlying information space. Adding to this, there is significant regional and institutional variance in the ways in which information of the chief complaints is expressed.

However, if we choose to record CCs as free text, how do we optimize the capture and understanding process? We want to computationally understand the CC so that we can automatically [4]

1. Detect a spike in the incidence of certain syndromes associated with bioterrorism and other emerging infections
2. Quickly group similar cases for research purposes
3. Expedite care by suggesting predefined care protocols to be initiated at triage
4. Improve patient safety by alerting providers to conditions they might not have considered
5. Improve efficiency by pre-selecting documentation templates for physicians

A key technical goal in providing a solution to these needs is to have a way to computationally reduce variety in the free-text CCs in a way that features with identical semantics are reduced to single descriptors (i.e., normalization). Optimally, and in relation to this, we should be able to extend normalization capability during capture. The end-user should be able to provide new normalization rules in an on-line training scenario.

2.2 The Chief Complaint Language

How can we efficiently normalize features in the CCs? First, let's consider the general characteristics of the English language typically used to record CCs in an US emergency department. The language of the CC reflects all of the challenging characteristics discussed in the introduction. The vocabulary is distinct even from that covering typical medical terminology. The use of abbreviations with non-standard expansions is commonplace. Phrase structure is highly compressed. Finally, the text is non-grammatical in terms of regular English, consisting mostly of irregular noun phrase fragments. Misspellings and accidental concatenations are also very frequent.

Because of these considerations, the use of standard analytical English parsers for structural normalization and standard semantic lexicons for lexical normalization are effectively ruled out. In effect, the chief complaints represent a language distinct from standard English. Thus, the remaining feasible approach is to develop an analytical model for a chief complaint language and to implement a normalization algorithm and a custom semantic lexicon based on this model. Purely statistical tools can't provide the high-granularity understanding required here.

First, like all natural human languages, the chief complaint language has compositional semantic features. To give a simple example, the term 'OD' has the meaning 'RIGHT EYE' (from 'ocular dextra') in the local context 'PAIN OD' and the meaning 'OVERDOSE' in the local context 'TYLENOL OD'. Thus, we want the normalizer to replace the term 'OD' with different normalizations depending on the local context. To achieve this, we chose to use a recursively applied rewriting system to model the grammar of the language. The production rules of the rewriting system implement the normalization.

Second, many grammar features are parameterizable. Thus, we included support for the use of regular expressions in the production rules to provide for their generalization. One example of such a production rule is the following:

TOOK\ *[0-9]*\ * MED → OVERDOSE (1)

This production rule also nicely illustrates how we can take advantage of the constraints provided by domain-specificity in normalization. In the context of the chief complaint language, it is a rational assumption that the appearance of the concept 'medication intake' has the specific meaning of an overdose as the patient's problem. This specific assumption, of course, is not generally valid in a less constrained setting of language use.

Third, there is much typographical variance in the chief complaints due to misspellings and accidental concatenations. To account for this, we included functionality to apply production rules in (conservatively) approximate matches.

Fourth, considering the listed application areas of normalized chief complaints, we want to assign normalized concepts with semantic category labels:

- (A) Anatomic location or function: Digit, Appendix, Urination, PICC-Line, etc.
- (C) Condition: Abrasion, Hypertension, Shingles, Paronychia, etc.
- (S) Symptom: Fever, Epistaxis, Facial-Droop, Hematemesis, etc.
- (P) Procedure: Evaluation, Lab-Tests, Placement, Nephrostomy, etc.
- (Q) Qualifier: Worse, Xanex, White, Under, Tree, Subtherapuitc, etc.
- (I) Ignore: Monday, Seen, Something, Still, This, Way, Yes, Wife, etc.
- (U) Unknown

2.3 Implementation

As shown in Fig. 1, the implemented normalization algorithm consists of three core steps: pre-processing, synonym matching and category assignment/training definition.

The resulting normalized chief complaints can be utilized for a number of purposes (including syndromic surveillance, suggestion of pre-defined care protocols at triage etc. as noted earlier). A key capability of the resulting system is that normalization capability can be easily increased in an on-line setting, potentially even integrating training to the initial CC entry at the ED. This would provide a facility to organically increase normalization capability on an on-demand basis.

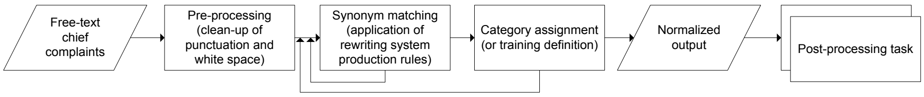

Free-text chief complaints
Pre-processing (clean-up of punctuation and white space)
Synonym matching (application of rewriting system production rules)
Category assignment (or training definition)
Normalized output
Post-processing task

Fig. 1. The normalization algorithm. Note that the synonym matching and category assignment steps are carried out recursively so that matches can be re-evaluated after any synonym replacements are made or new training is added.

An initial version of the algorithm and its evaluation were presented in [4]. An enhanced version and its evaluation are provided in [5], along with details on the physical architecture of the normalization system.

2.4 Evaluation

Prior evaluations of this normalization algorithm ([4], [5]) showed that it was highly accurate in terms of completeness and correctness when the training set and test set were from the same institution. Good completeness performance was also observed with a training set composed from multiple hospitals [6]. The purpose here is to evaluate correctness performance with a multi-hospital training set and also to determine whether approximate matching increases normalization completeness without degrading correctness with this training set.

A dataset containing 25,000 CC's (5,000 each from 5 Boston-area hospitals) was randomly divided into 2 sets: a training set and a test set. Narrative style and linguistic complexity of the CCs varied considerably between hospitals. Normalizer training was performed by a graduate student with emergency medicine and bioinformatics experience (MA) with assistance and oversight from (SN) and (LN). The training took approximately 36 working hours to complete. Once training was complete, the test set was unblinded and normalized. The percentage of completely normalized CCs (those results which contained no assignments to the category “unknown”) was then calculated as well as the percentage of correctly normalized CCs (those results which contained erroneous category assignments). See Table 1 for the results.

The McNemar test was performed to test whether approximate matching has an effect on the percentage of completely and correctly normalized CCs. Since the value of chi-square statistic with 1 degree of freedom was 54.598 ($p<.0001$) for total completeness data, we could conclude that approximate matching significantly increased the percentage of completely normalized CCs [6]. Similarly, as the value of chi-square statistic was 1.333 for total correctness data, we could conclude that approximate matching didn’t significantly degrade the percentage of correctly normalized CCs.

Table 1. Percentages of completely and correctly normalized CCs (95% CI). Completeness results were originally published in [6]. (1): without approximate matching, (2): with.

Hospital	% Completely Normalized (1), N=2500	% Correctly Normalized (1), N=100	% Completely Normalized (2), N=2500	% Correctly Normalized (2), N=100
A	90.2 [89.0-91.3]	99.0 [94.6-99.8]	93.0 [92.0-94.0]	99.0 [94.6-99.8]
B	86.4 [85.0-87.7]	99.0 [94.6-99.8]	89.4 [88.1-90.5]	99.0 [94.6-99.8]
C	97.6 [97.0-98.1]	100 [96.3-100]	97.9 [97.3-98.4]	100 [96.3-100]
D	75.7 [74.0-77.4]	99.0 [94.6-99.8]	80.8 [79.2-82.4]	98.0 [93.0-99.5]
E	79.5 [77.9-81.0]	98.0 [93.0-99.5]	83.9 [82.4-85.3]	96.0 [90.2-98.4]
Total	86.0 [85.4-86.6]	99.0 [97.7-99.6]	89.1 [88.5-89.6]	98.4 [96.9-99.2]

3 Discussion

As shown by the presented case study, domain-specific language modeling implemented as a case-specific rewriting system is a highly promising tool for the computational understanding of a specific class of medical text. Further work is required to establish whether normalization performance remains constant when the training is carried out by different persons. Future research directions include the application of similar methods to other non-standard medical texts.

References

1. Taira, R., Bashyam, V., Kangarloo, H.: A Field Theoretical Approach to Medical Natural Language Processing. IEEE Trans. Inf. Technol. Biomed., vol. 11, no. 4, pp. 364-75 (2007)
2. National Center for Injury Prevention and Control: Data elements for emergency department systems, release 1.0. Atlanta, GA (1997)
3. Thompson, D., Eitel, D., Fernandes, C., Pines, J., Amsterdam, J., Davidson, S.: Coded Chief Complaints--automated analysis of free-text complaints. Acad. Emerg. Med., vol. 13, no. 7, pp. 774-82 (2006)
4. Nathanson, L., Mandl, K., Olson, K., Ladapo, J., Shapiro, N.: Evaluation of an Algorithm for the Normalization of Chief Complaints. Acad. Emerg. Med., vol. 12, no. 5 suppl 1, pp. 101 (2005)
5. Niiranen, S., Yli-Hietanen, J., Nathanson, L.: Towards Reflective Management of Emergency Department Chief Complaint Information. IEEE Trans. Inf. Technol. Biomed. (in press)
6. Aswell, M., Niiranen, S., Nathanson, L.: Enhanced Normalization of Emergency Department Chief Complaints. In: American Medical Informatics Association 2007 Annual Symposium, November 10-14, Chicago, Illinois, USA (2007)

Automated Text Segmentation and Topic Labeling of Clinical Narratives

Hanna Suominen[1,2], Sampo Pyysalo[1,2], Filip Ginter[2], and Tapio Salakoski[1,2]

[1] Turku Centre for Computer Science (TUCS) and
[2] University of Turku, Department of Information Technology
Joukahaisenkatu 3-5 B, 20520 Turku, Finland
firstname.lastname@utu.fi

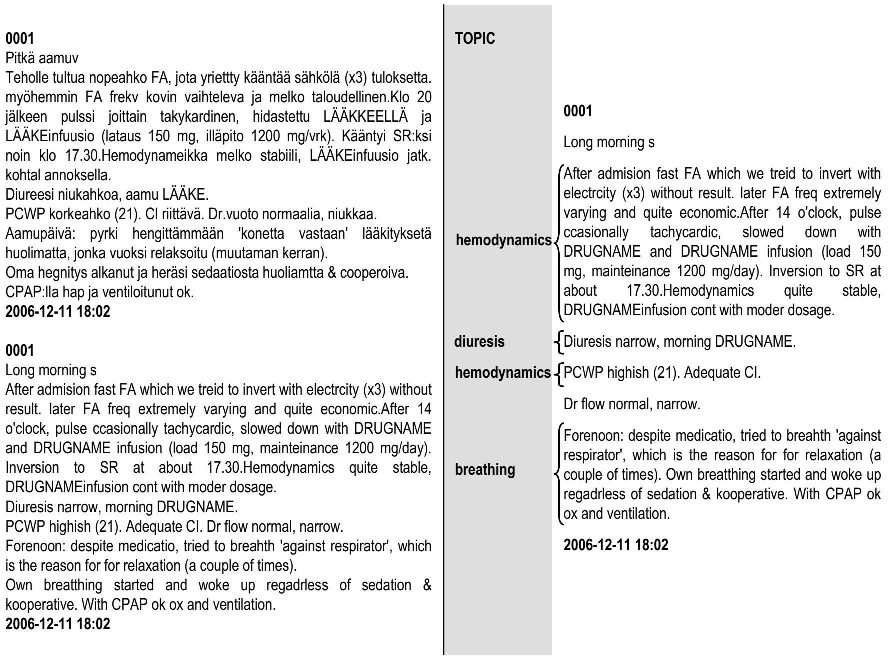
Abstract. Electronic patient information systems include numerous functionalities to support clinical judgment and decision-making, but their capabilities to analyze free-text narratives are limited. We apply Hidden Markov Models to divide Finnish intensive care nursing notes into topically coherent segments and assign a topic label to each segment. The method notably outperforms a keyword-based baseline already with a relatively small amount of training data. The result holds the promise of increased information search speed and a more comprehensive overall picture about patients.

1 Introduction

Modern electronic patient information systems include numerous functionalities to support clinical judgment and decision-making [1]. For example, they generate statistics, trends and alerts from the patient data. However, although a substantial amount of information is documented as free-text notes, referred to as narratives, automated processing is typically limited to the numerical or structured parts of the patient records. Text mining applications in clinical use, such as MedLEE [2] and Autocoder [3] for English text, are rare in particular for minority languages.

In this study, we present a text segmentation and topic labeling (TS & TL) method for dividing Finnish narratives automatically into topically coherent, non-overlapping segments (Figure 1). The resulting type of structure has been empirically shown to increase the information search speed of clinicians [4]. The domain we consider is intensive care (IC), as its complexity, information richness, and fast pace make decision-making particularly challenging.

In the clinical domain, TS techniques have previously been applied, for example, to temporal order analysis of medical discharge summaries [5]. The method solves the problem sequentially by using a statistical parser to segment the sentences into clauses, a classifier to predict the segment boundaries between the clauses and finally another classifier to decide for every segment pair their time-wise order. Another application related to this study is TS & TL of medical narratives from radiology and urology departments [6]. Although it contains



0001
Pitkä aamuv
Teholle tultua nopeahko FA, jota yriettty kääntää sähkölä (x3) tuloksetta. myöhemmin FA frekv kovin vaihteleva ja melko taloudellinen.Klo 20 jälkeen pulssi joittain takykardinen, hidastettu LÄÄKKEELLÄ ja LÄÄKEinfuusio (lataus 150 mg, illäpito 1200 mg/vrk). Kääntyi SR:ksi noin klo 17.30.Hemodynameikka melko stabiili, LÄÄKEinfuusio jatk. kohtal annoksella.
Diureesi niukahkoa, aamu LÄÄKE.
PCWP korkeahko (21). CI riittävä. Dr.vuoto normaalia, niukkaa.
Aamupäivä: pyrki hengittämmään 'konetta vastaan' lääkityksetä huolimatta, jonka vuoksi relaksoitu (muutaman kerran).
Oma hegnitys alkanut ja heräsi sedaatiosta huoliamtta & cooperoiva.
CPAP:lla hap ja ventiloitunut ok.
2006-12-11 18:02

0001
Long morning s
After admision fast FA which we treid to invert with electrcity (x3) without result. later FA freq extremely varying and quite economic.After 14 o'clock, pulse ccasionally tachycardic, slowed down with DRUGNAME and DRUGNAME infusion (load 150 mg, mainteinance 1200 mg/day). Inversion to SR at about 17.30.Hemodynamics quite stable, DRUGNAMEinfusion cont with moder dosage.
Diuresis narrow, morning DRUGNAME.
PCWP highish (21). Adequate CI. Dr flow normal, narrow.
Forenoon: despite medicatio, tried to breahth 'against respirator', which is the reason for for relaxation (a couple of times).
Own breatthing started and woke up regadrless of sedation & kooperative. With CPAP ok ox and ventilation.
2006-12-11 18:02

TOPIC

0001

Long morning s

hemodynamics After admision fast FA which we treid to invert with electrcity (x3) without result. later FA freq extremely varying and quite economic.After 14 o'clock, pulse ccasionally tachycardic, slowed down with DRUGNAME and DRUGNAME infusion (load 150 mg, mainteinance 1200 mg/day). Inversion to SR at about 17.30.Hemodynamics quite stable, DRUGNAMEinfusion cont with moder dosage.

diuresis Diuresis narrow, morning DRUGNAME.

hemodynamics PCWP highish (21). Adequate CI.

Dr flow normal, narrow.

breathing Forenoon: despite medicatio, tried to breahth 'against respirator', which is the reason for for relaxation (a couple of times). Own breatthing started and woke up regadrless of sedation & kooperative. With CPAP ok ox and ventilation.

2006-12-11 18:02

Fig. 1. An anonymized illustration of the Finnish data accompanied with its English translation preserving typographical errors and an example segmentation into topics.

a classifier based on the order of sections and other statistical features of the training data, it mainly relies on hard-coded headlining rules, linguistic cues and lexical patterns seen within training examples. Finally, a system classifying segments of IC patient narratives with respect to the topics of *breathing*, *blood circulation* and *pain* has been introduced [7]; it does not, however, perform automated TS at all.

2 Patient data and its linguistic processing

Anonymized nursing narratives[3] of 516 adult IC patients were used in this study. They covered the whole in-patient time and were written mainly for intra-unit information exchange. We chose these notes because their use in direct care is hindered due to the large quantity of text.

The data set included altogether 17140 patient and nursing shift-specific documents, which we call shifts (Figure 1). Each shift contained, on average, 73 tokens (including punctuation). The vocabulary was highly specialized, with a

[3] Collected retrospectively from January 1, 2005 to August 1, 2006 with proper permissions (Statutes of Finland: Medical research act 488/1999 and decree 986/1999) for the Louhi project (www.med.utu.fi/hoitotiede/tutkimus/tutkimusprojektit/louhi).

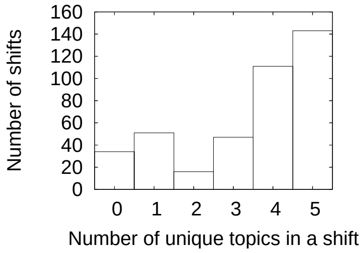


Fig. 2. A histogram illustrating the number of topics discussed per shift.

substantial amount of unit-specific practices and domain terminology. Approximately half of the shifts were structured by using colon-separated, but non-standardized, headings, as *Hemodynamics*, *HAEMODYNAMICS*, *H e m o d*, and *Homedynamics*. The most common documentation topics were *breathing*, *hemodynamics*, *consciousness*, *relatives*, and *diuresis*. We selected these to be used as TS topics due to their prevalence in the narratives; by recognizing them automatically, building an overall picture about their development in time could be supported, for example, through topical highlighting.

To create topic-annotated data for experiments, we randomly chose three shifts per patient from the records of 135 patients and tagged text segments relevant to the five most common topics by using the Knowtator tool [8] of Protégé 3.3.1 Ontology Editor and Knowledge Acquisition System[4]. Irrelevant parts were given the label *other*. With problematic phrases, an IC nursing specialist was consulted. The average shift length was 78 tokens while the average segment length was only 18 tokens. Typically, all or almost all five topics were discussed within one shift although 34 shifts contained none of the topics (Figure 2).

To reduce data sparseness caused by the highly inflective nature of Finnish, we lemmatized the data using a version of the FinTWOL Finnish morphological analyser[5] [9] whose lexicon was extended by approximately 3500 clinical domain terms. For every word analyzed by FinTWOL, we used the first lemma given.

3 Method and its performance evaluation setting

Let us denote the topics of interest as $q_i, i \in \{1, \ldots, N_q\}$. Our TS & TL task is to infer for the input word sequence $w = [w(1) \ldots w(T)]$ the topic sequence $q = [q(1) \ldots q(T)]$, where $w(t)$ belongs to the vocabulary $\{w_1, \ldots, w_{N_w}\}$ of N_w unique words and $q(t) \in \{q_1, \ldots, q_{N_q}\}$ for all $t \in \{1, \ldots, T\}$. A convenient way to model this sequence labeling problem is to use a first-order Hidden Markov Model (HMM) (see, e.g., [10]), where a particular hidden variable $q(t)$ only depends on the previous hidden state $q(t-1)$, an observed variable $w(t)$ is only dependent on the value of the hidden variable $q(t)$, and the random variable

[4] http://protege.stanford.edu/

[5] http://www.lingsoft.fi/

describing the start of the chain is uniformly distributed. Formally, if $\mathcal{Q}$ is the space of all hidden state sequences, we infer the best q by solving

$$\arg\max_{q\in\mathcal{Q}} P(w(1)|q(1)) \prod_{t=2}^{T} P(w(t)|q(t))P(q(t)|q(t-1)).$$

We trained the HMM with approximately half of the annotated shifts and tested it with the other half. No patient record was divided between the two sets. The smoothing model and its parameter were selected on the training set by a separate search of the parameter space so as to avoid over-fitting the test set. The selected optimal model was Lidstone (add-γ) smoothing (see, e.g., [11, p. 204]) with $\gamma = 0.3$.

The baseline algorithm implements a simple topic keyword search: We first searched for the five topic keywords. Then, we assigned each word to a labeled segment corresponding to the previous seen topic until the end of the shift. We gave the assigned label at the start of each shift the initial value of *other*. This baseline was chosen because it inherently resembles the documentation structure. To allow the baseline to benefit from the normalizing effect of morphological analysis, TS & TL was performed with the data processed with FinTWOL. In evaluation, we measured the token-wise average TL accuracy over the whole test data.

4 Results and conclusion

HMM performing TS & TL of IC nursing notes with the lemmatized text notably outperformed the keyword-based baseline of 66.95% already with as few as 2000 words of training data (Figure 3). This corresponds approximately to tagging 20 shifts like the one given in Figure 1. The performance increase can be seen to level off after about 8000 words. Linguistic processing contributed to the performance, but its significance diminished by increasing the amount of training data: the accuracy of HMM with the lemmatized data was 82.93%, whereas the respective number without lemmatization was 81.22%.

Our results hold promise for improving the functionality of electronic patient information systems: the method is easy to implement and its integration should be relatively straightforward. Highlighting of the most prevalent topics is likely to expedite information search and offer improved capabilities to build an overall picture about their development in time. In order to allow freely chosen segmentation topics, we have also developed an unsupervised method for the task [12]. Future work will include a pilot study testing our methods in clinical use. Other interesting research directions are generating trends and summarizing text on the basis of the automatically topic-labeled narratives.

Acknowledgments

We gratefully acknowledge the financial support of the Academy of Finland and the Finnish Funding Agency for Technology and Innovation, Tekes. We also express our

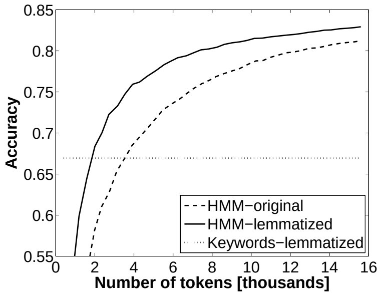


Fig. 3. Learning curves for the HMMs with and without linguistic processing and the overall accuracy of the keyword-based baseline.

gratitude to Heljä Lundgrén-Laine for help in annotation, to Philip Ogren for assistance with Knowtator and to Sari Ahonen and Simo Vihjanen from Lingsoft Inc. for extending FinTWOL.

References

1. Hanson, C., Marshall, B.: Artificial intelligence applications in the intensive care unit. Crit Care Med **29**(2) (2001) 427–435
2. Mendonça, E., Haas, J., Shagina, L., Larson, E., Friedman, C.: Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform **38**(4) (2005) 314–321
3. Pakhomov, S., Buntrock, J., Chute, C.: Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc **13**(5) (2007) 516–525
4. Tange, H., Schouten, H., Kester, A., Hasman, A.: The granularity of medical narratives and its effect on the speed and completeness of information retrieval. J Am Med Inform Assoc **5**(6) (1998) 571–582
5. Bramsen, P., Deshpande, P., Lee, Y., Barzilay, R.: Finding temporal order in discharge summaries. AMIA Annu Symp Proc (2006) 81–85
6. Cho, P., Taira, R., Kangarloo, H.: Automatic section segmentation of medical reports. AMIA Annu Symp Proc (2003) 155–159
7. Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., Salakoski, T.: Towards automated classification of intensive care nursing narratives. Int J Med Inform **76**(S3) (2007) S362–S368
8. Ogren, P.: Knowtator: A Protégé plug-in for annotated corpus construction. In: Proc HLT-NAACL 2006, Morristown, NJ, USA, ACL (2006) 273–275
9. Koskenniemi, K.: Two-level model for morphological analysis. In: Proc IJCAI 83. Volume 2., Karlsruhe, Germany, Morgan Kaufmann (1983) 683–685
10. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2) (1989) 257–286
11. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, USA (1999)
12. Ginter, F., Suominen, H., Pyysalo, S., Salakoski, T.: Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. In: Proceedings of SMBM'08. (2008) To appear.

LOUHI'08

The First Conference on Text and Data Mining of Clinical Documents

Work in Progress Proposals (WOPPs)

Louhi'08 gives room to a new innovative type of submission: work in progress proposals (WOPPs). A WOPP is a peer-reviewed extended abstract that expresses compelling ideas, proposes new practical or methodological challenges, unfurls innovative research questions, showcases or demands novel initiatives (e.g. joint projects), etc.

The presenters are given room for a short oral presentation in the plenum followed by room for open discussion. They receive the unique opportunity to challenge the plenum with novel ideas and to gather views on the prospects of the ongoing work.

Feasibility of a Computer-Based Decision Support System – a Study Plan with Primary Care Nurses

Tiina Kortteisto[1], Ilkka Kunnamo[2], and Minna Kaila[1,3]

[1] Tampere School of Public Health, FI-33014 University of Tampere, Finland
[2] Duodecim Medical Publications Ltd. P.O. Box 874, FI-00101 Helsinki, Finland
[3] Finohta / STAKES Finn-Medi 3, Biokatu 10, FI-33520 Tampere, Finland

Abstract. The study will explore the feasibility of a computer-based decision support system (CDSS) in primary care nurses' practices for diabetes and dialysis patients in one rural health centre. The target CDSS has been developed in the Evidence-Based Medicine electronic Decision Support (EBMeDS) project and integrated into an electronic diabetes management system. Data will be gathered from four sources: 1) log file information, 2) nurses' evaluations of decision support reminders, 3) a survey, and 4) focus group interviews of nurses. We will analyse both positive and negative consequences of the use of computer-based decision support.

1 Introduction

The definition and content of computer-based decision support systems (CDSSs) vary with the experts and contexts involved [1, 2, 3]. There are three basic principles for developing a CDSS in the Finnish Evidence-Based Medicine electronic Decision Support (EBMeDS) project: 1) automatic provision of computer-based decision support (CDS) as part of the clinician's work flow, 2) provision of evidence-based recommendations rather than assessments only, and 3) provision of decision support at the time and place of decision-making; these were derived from systematic reviews [4, 5]. At present, the EBMeDS system is integrated into an electronic diabetes management system, Prowellness (PW), which provides three levels of reminders (I, II, III), categorised by importance for a specific patient. The reminder can be triggered by the structured patient data, diagnosis, medication, or laboratory results when a nurse opens the PW system. More information on the EBMeDS system is available from the project's Web pages (accessible via http://www.kaypahoito.fi/decisionsupport/decisionsupport.htm).

Experience and evidence of CDSS in the primary care nurse context is scarce [6, 7]; therefore, this is the setting for our study.

2 Study Questions

The aim of the study is to explore the feasibility of the use of CDS in primary care nurses' care for diabetes and dialysis patients. The specific study questions will be 1) what reminders are triggered in target patient groups and how often (via log file data), 2) whether the triggered patient-specific reminders are accepted and exploited by the nurses (as evidenced by patient document and survey material), and 3) what the nurses' experiences are of use of CDS reminders (as shown by focus group interviews).

3 Method

Data will be gathered from four sources, via different methods. 1) The log file of the PW system will automatically gather information on the number and type of reminders triggered, without any patient-specific information. 2) Printed patient documents containing nurses' evaluations are to be gathered in two ways. First, during encounters, CDS reminders are triggered in the PW system when the nurse opens the patient record and enters data. The nurse will print the reminder and answer three questions: 1) was the reminder justified ('yes' or 'no'); 2) did he or she comply with the recommendation ('yes' or 'no'); and 3) if not, why not? Second, a virtual health check for specified patients will be performed. This means that all reminders are triggered (all available scripts are executed) as a batch run for a selected group of the nurse's patients. The nurse receives a list of reminders for each individual patient and evaluates the reminders again by using the three questions above. Using the virtual health check enables us to gather data also for those patients who do not visit the nurse during the study period. 3) The nurses' evaluation of the feasibility of every reminder is to be gathered via a structured questionnaire. The specific survey questions are now being developed. 4) After data collection is complete, we will organise focus group meetings to gather qualitative data of nurses' experiences of the use of CDS reminders, and of the influence of the reminders on their actions. A structured agenda will be developed for these discussions.

4 Ethical Questions

The study protocol has been accepted by the local ethics committee, and the chief physician of the health centre has authorised the study. The participating nurses have received oral and written information, and each has given informed consent. This study focuses on nurses' behaviour and decision-making, with anonymous patient data analysed only in the context of the CDS reminders. Both positive and negative consequences of the use of CDS will be assessed.

5 Discussion

This is a typical real-world study, wherein the researchers are constrained by practicalities. Several, even important, features have been excluded for practical reasons. For instance, it would be interesting to target physicians as well as nurses, but there is a severe shortage of physicians at the study health centre, making it unfeasible to place any extra burden on those who remain. Also, social and contextual factors may affect the use of CDS among primary care nurses [1, 2, 7, 8], which is why we plan to use both quantitative and qualitative study methods. We target the management of diabetes, which is undoubtedly the most important chronic disease in Finland and is predicted to consume a large proportion of the available health care resources [9]. Nurses already contribute significantly to the care of this patient group. However, more evidence would be needed of technological advance that improves the processes for, and quality of, care of large number of patients in the future [7, 10].

At present, the use of CDS is hindered by the lack of structured (coded) patient data. For example, the diagnoses are recorded only as free text in the nurses' records. Computerised tools for free text analysis could facilitate recording of data in structured format.

References

1. Berlin, A., Sorani, M., Sim, I.: A Taxonomic Description of Computer-Based Clinical Decision Support Systems. J. Biomed. Inform. 39, 656--667 (2005)
2. Delpierre, C., Cuzin, L., Fillaux, J., Alvarez, M., Massip, P., Lang, T.: A Systematic Review of Computer-Based Patient Record Systems and Quality of Care: More Randomized Clinical Trials or a Broader Approach? Int. J. Qual. Health Care 5, vol. 16, 407--416 (2004)
3. Niès, J., Colombet, I., Degoulet, P., Durieux, P.: Determinants of Success for Computerized Clinical Decision Support Systems Integrated into CPOE Systems: a Systematic Review. In: AMIA Symposium Proceedings, pp. 594--598 (2006)
4. Garg, A.X., Adhikari, N.K.J., McDonald, H., Rosas-Arellano, M.P., Devereaux, P.J., Beyene, J., Sam, J., Haynes, R.B.: Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes. JAMA 10, vol. 293, 1223--1238 (2005)
5. Kawamoto, K., Houlihan, C.A., Balas, E.A., Lobach, D.F.: Improving Clinical Practice Using Clinical Decision Support Systems: a Systematic Review of Trials to Identify Features Critical to Success. BMJ 330, 765--772 (2005)
6. Varonen, H., Kaila, M., Kunnamo, I., Komulainen, J., Mäntyranta, T. Tietokoneavusteisen Päätöksentuen Avulla kohti Neuvovaa Potilaskertomusta. Duodecim 122, 1174--1181 (2006)
7. Randell, R., Mitchell, N., Dowding, D., Cullum, N., Thompson, C.: Effects of Computerized Decision Support Systems on Nursing Performance and Patient Outcomes: a Systematic Review. J. Health Serv. Res. Policy 4, vol. 12, 242--249 (2007)
8. Kaplan, B.: Evaluating Informatics Applications – Clinical Decision Support Systems Literature Review. Int. J. Med. Inform. 64, 15--37 (2001)
9. Winell, K., Reunanen, A.: Diabetes Barometer 2005. Finnish Diabetes Association, Tampere (2006)
10. Eccles, M.P., Whitty, P.M., Speed, C., Steen, I.N., Vanoli, A., Hawthorne, G.C., Grimshaw, J.M., Wood, L.J., McDowell, D. : A Pragmatic Cluster Randomised Controlled Trial of a Diabetes Recall and Management System : the DREAM Trial. Implement Sci. 2, 6 (2007)

Diagnosing Diagnoses in Swedish Clinical Records

Sumithra Velupillai, Hercules Dalianis and Martin Hassel

DSV, KTH-Stockholm University, 164 40 Kista, Sweden
{sumithra,hercules,xmartin}@dsv.su.se

1 Introduction

Electronic clinical record systems are becoming the standard for many hospitals, providing an extensive amount of valuable information which could be used for important research in different research areas. In our project, we have access to a large set of de-identified clinical records from several departments in one of the largest hospitals in Sweden: Karolinska University Hospital. To our knowledge, this set is unique in at least two ways; it is the first set of clinical records written in Swedish, and it is the first set covering several medical departments, thus providing an invaluable data set for many research areas.

Clinical records contain both structured and unstructured entries, such as measurement values and sections of free text. However, the free text sections of clinical records have not, until recently, been used for further research. Such sections hold great potential for inventive text mining and computational linguistics research.

The language use in clinical records is very specific and noisy, containing domain-specific vocabulary, and often ad-hoc abbreviations and misspellings. Moreover, these types of text contain a potentially large amount of speculation, uncertainty and negation together with certainty and confirmation. This property is significant for the diagnosis and documentation procedure, and is very important to extract. For many text mining and information extraction tools, such issues are seldom taken into account, which we believe is problematic. These aspects have gained a lot of interest recently, and many methods for handling such parts in text sets have been proposed. However, most experiments have been performed on text sets in English, and mostly on similar contents. We plan to apply and evaluate existing state-of-the-art methods on Swedish clinical records. Moreover, we plan to develop these methods further with the goal of being as language independent as possible and generic for different medical specializations.

2 Related Work

Research on speculative language, or identification of both negations, uncertainties and other hedging cues in text has, in the (bio)medical domain, been performed both on biomedical scientific literature (full articles as well as abstracts) and clinical records. Many methods have been developed using handcrafted rule-based negation and uncertainty detection modules (see for instance Szarvas et al. (2008)). In Kilicoglu & Bergler (2008), the hedge classification dataset developed by Medlock & Briscoe (2007) was used, utilizing existing lexical resources with an extension of syntactic patterns and weighting schemes. In Szarvas et al. (2008), the creation of the BioScope corpus is described, a project with the aim of creating an annotated text set which can be used for developing and evaluating automatic classification systems for this specific phenomenon. The created corpus consists of biomedical scientific full papers and abstracts as well as medical free texts. A corpus consisting only of clinical free text has been used in a shared task on multi-label classification described in Pestian et al. (2007).

3 Proposed Work

The set of Swedish clinical records that we have access to is, as stated above, unique in many ways. The risk for accessing information that may be used to identify patients is an important aspect that has to be taken into account. The records we have obtained have been automatically de-identified, but many records may still have information in the free-text sections that may be used for identification, such as phone numbers, family member names, specific occupations etc. We intend to use and evaluate de-identification methods such as named-entity recognizers, in order to remove the risk of accessing private data. For the project described here, we propose to extract a small (fully de-identified) balanced subset for annotation of negation, speculation and certainty, based on the guidelines described in Szarvas et al. (2008). This set of annotated data will be classified applying current state-of-the-art methods, such as the ones described in Kilicoglu & Bergler (2008), and evaluated on our data set, which differs both in language and in text type.

From a small amount of clinical records from the Rheumatology clinic at the Karolinska University Hospital, we have identified several examples of both uncertain and certain diagnoses:

(1) Patient med oklar myalgi, muskelsvaghet med 10 mg Prednisolon, har ingen CK stegring, inga säkra förändringar på muskelbiopsi. Negativ EMG. Oklar diagnos. Statinutlöst myopati?
Patient with unclear myalgia, muscle weakness with 10 mg Prednisolon, no CK increase, no certain changes in muscle biopsy. Negative EMG. Unclear diagnosis. Statin triggered myopathia?

(2) Tydlig effekt av Methotrexate o Remicade i händerna, dock resterande sjd-aktivitet. Kan absolut inte avstå från NSAID.
Clear effect of Methotrexate and (abbr) Remicade in the hands, yet remaining signs of active disease (abbr). Still in need of NSAID (formulated with negation in Swedish).

As we can see in the examples above, conclusions (uncertain or certain) may span over sentence boundaries. Therefore, we will extend the annotations to span over sequences of sentences covering a diagnosis. In these guidelines, speculative elements are marked by angled brackets (<>), and negative elements are marked by square brackets ([]). We will also extend the guidelines to annotate elements that indicate certainty, with curly brackets ({}). The certain diagnosis in Example 2 for instance, would be annotated the following way:

(3) ({Tydlig} effekt av Methotrexate o Remicade i händerna), dock resterande sjd-aktivitet. Kan absolut ([inte] avstå från NSAID).

We propose the following work plan:

- Annotate a (fully de-identified) subset of the data set
- Apply existing state-of-the-art tools for classification of speculative sequences
- Analyze and evaluate the results, especially with regards to differences in the following aspects: language, medical specialization, and style and tradition in writing clinical records
- Develop methods for improving performance, primarily using word space models (and possibly extensions to constructions that can be modeled as words)

As many natural language processing tools will be needed in preprocessing steps, especially the ones used for de-identifying the full data set, current tools may not work optimally for these types of texts in Swedish. Evaluation and fine-tuning of such preprocessing steps will also have to be made. However, by using word space models, thus utlizing distributional patterns and relations in the text sets, heavy lexical and linguistic resources will not be needed once the records are de-identified.

Acknowledgements

Our thanks go to Vinnova (Swedish Governmental Agency for Innovation Systems) for funding the KEA-project (grant p29670-1), and to Jussi Karlgren (SICS) for initializing the idea and commenting the proposal.

References

H. Kilicoglu and S. Bergler: Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In BioNLP 2008: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, Association for Computational Linguistics, 46–53, June 2008.

B. Medlock and T. Briscoe: Weakly Supervised Learning for Hedge Classification in Scientific Literature. In Proceedings of the 45th Meeting of the Association for Computational Linguistics, 647–656, Prague, Czech Republic, June 2007.

J. P. Pestian, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K. B. Cohen and W. Dutch: A Shared Task Involving Multi-label Classification of Clinical Free Text. In Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, Association for Computational Linguistics, 97–104, June 2007.

G. Szarvas, V. Vincze, R. Farkas and J. Csirik: The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Texts. In BioNLP 2008: Current Trends in Biomedical Natural Language Processing, Columbus, Ohio, Association for Computational Linguistics, 38–45, June 2008.

Turku Centre for Computer Science
TUCS General Publications

30. **Mats Aspnäs, Christel Donner, Monika Eklund, Ulrika Gustafsson, Timo Järvi and Nina Kivinen (Eds.)**, Turku Centre for Computer Science, Annual Report 2003

31. **Andrei Sabelfeld (Editor)**, Foundations of Computer Security

32. **Eugen Czeizler and Jarkko Kari (Eds.)**, Proceedings of the Workshop on Discrete Models for Complex Systems

33. **Peter Selinger (Editor)**, Proceedings of the 2nd International Workshop on Quantum Programming Languages

34. **Kai Koskimies, Johan Lilius, Ivan Porres and Kasper Østerbye (Eds.)**, Proceedings of the 11th Nordic Workshop on Programming and Software Development Tools and Techniques, NWPER'2004

35. **Kai Koskimies, Ludwik Kuzniarz, Johan Lilius and Ivan Porres (Eds.)**, Proceedings of the 2nd Nordic Workshop on the Unified Modeling Language, NWUML'2004

36. **Franca Cantoni and Hannu Salmela (Eds.)**, Proceedings of the Finnish-Italian Workshop on Information Systems, FIWIS 2004

37. **Ralph-Johan Back and Kaisa Sere**, CREST Progress Report 2002-2003

38. **Mats Aspnäs, Christel Donner, Monika Eklund, Ulrika Gustafsson, Timo Järvi and Nina Kivinen (Eds.)**, Turku Centre for Computer Science, Annual Report 2004

39. **Johan Lilius, Ricardo J. Machado, Dragos Truscan and João M. Fernandes (Eds.)**, Proceedings of MOMPES'05, 2nd International Workshop on Model-Based Methodologies for Pervasive and Embedded Software

40. **Ralph-Johan Back, Kaisa Sere and Luigia Petre**, CREST Progress Report 2004-2005

41. **Tapio Salakoski, Tomi Mäntylä and Mikko Laakso (Eds.)**, Koli Calling 2005 - Proceedings of the Fifth Koli Calling Conference on Computer Science Education

42. **Petri Paju, Nina Kivinen, Timo Järvi and Jouko Ruissalo (Eds.)**, History of Nordic Computing - HiNC2

43. **Tero Harju and Juhani Karhumäki (Eds.)**, Proceedings of the Workshop on Fibonacci Words 2006

44. **Michal Kunc and Alexander Okhotin (Eds.)**, Theory and Applications of Language Equations, Proceedings of the 1st International Workshop, Turku, Finland, 2 July 2007

45. **Mika Hirvensalo, Vesa Halava and Igor Potapov, Jarkko Kari (Eds.)**, Proceedings of the Satellite Workshops of DLT 2007

46. **Anne-Maria Ernvall-Hytönen, Matti, Jutila, Juhani Karhumäki and Arto Lepistö (Eds.)**, Proceedings of Conference on Algorithmic Number Theory 2007

47. **Ralph-Johan Back and Ion Petre (Eds.)**, Proceedings of COMPMOD 2008

48. **Elena Troubitsyna (Editor)**, Proceedings of Doctoral Symposium held in conjunction with Formal Methods 2008

49. **Reima Suomi and Sanna Apiainen (Eds.)**, Promoting Health in Urban Living: Proceedings of the Second International Conference on Well-being in the Information Society (WIS 2008)

50. **Aulis Tuominen, Jussi Kantola, Arho Suominen and Sami Hyrynsalmi (Eds.)**, NEXT 2008 - Proceedings of the Fifth International New Exploratory Technologies Conference

51. **Tapio Salakoski, Dietrich Rebholz-Schuhmann and Sampo Pyysalo (Eds.)**, Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)

52. **Helena Karsten, Barbro Back, Tapio Salakoski, Sanna Salanterä and Hanna Suominen (Eds.)**, The Proceedings of the First Conference on Text and Data Mining of Clinical Documents (Louhi’08)

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Information Technologies

Turku School of Economics

- Institute of Information Systems Sciences

ISBN 978-952-12-2129-3
ISSN 1239-1905

Proceedings of Louhi’08

Proceedings of Louhi’08

Proceedings of Louhi’08