

**MEETING OF WHO COLLABORATING CENTRES
FOR THE FAMILY OF INTERNATIONAL CLASSIFICATIONS**

Tunis, Tunisia
29 Oct. - 4 Nov. 2006

**Language Standardization for Mortality Coding
A German Approach**
Stefanie Weber, Orlando Özer

Abstract

In Germany 16 counties ("Länder") are manually coding their mortality data separately, and for a long time there has been the urge to harmonize the regional differences in coding. Even though regular education was conducted only up to 70 % of the data seem to be coded manually in the same way. Reasons for the difference in assigning the underlying cause vary from unequal knowledge levels to different interpretation of the instruction rules and medical expressions.

Automated coding was considered a good way of reaching unified coding results but existing systems do not handle the German language. As the development of a separate tool for Germany was considered too resource-intensive and as the translation of the American dictionary of MMDS too error-prone, the development of IRIS was watched closely and Germany decided to try to use IRIS for mortality coding in Germany.

As IRIS does not only offer automated coding with the entry of ICD-10 codes but as well allows to include a dictionary and standardization tables for a specific language, we decided to adapt the German morbidity index to mortality needs and to create language standardization tables specific to the German mortality vocabulary.

The first tests showed that the vocabulary of the morbidity index (74,000 terms) does not suffice for use in automated coding. Analysis of terms from death certificates showed the great variety of text combination and the creativity of German physicians towards abbreviations and new terms.

Therefore standardization through regular expressions, as used in IRIS, was enforced. Even though the results show the obstacles and limitations of standardization for a complicated language as German they are quite encouraging. In a relatively short period of time the single line recognition of texts could be raised from 30% to 70%. Therefore we are quite confident that for the planned implementation in Germany for 2008 the system will be very efficient and will harmonize the output data significantly.

This document is not issued to the general public, and all rights are reserved by the World Health Organization (WHO). The document may not be reviewed, abstracted, quoted, reproduced or translated, in part or in whole, without the prior written permission of WHO. No part of this document may be stored in a retrieval system or transmitted in any form or by any means - electronic, mechanical or other - without the prior written permission of WHO.

The views expressed in documents by named authors are solely the responsibility of those authors.

Content

Abstract 1

Introduction 3

Language Standardization for Mortality coding 3

 First results 3

 Progress in testing 4

 Analysis 5

Discussion 6

Introduction

In many spoken languages changes and amendments to words occur over the years. Grammatical flections are used and composed words can be altered through rearrangement of its components. New vocabulary is added to a language whereas some old terms disappear from the used vocabulary.

While this phenomenon is a positive aspect of a spoken language, it is a complication to electronic processing of the vocabulary.

Lots of research has already been invested in this topic and many approaches towards natural language processing have been created. Most of these approaches use "nearest match" algorithms or phrase interpretation.

For the automated processing of the vocabulary on death certificates these approaches are insufficient. The entries are short and the alteration of only one letter can change the meaning of the word used to express mortality information.

Language Standardization for Mortality coding

As Germany would like to implement automated mortality coding and decided to use IRIS for this project, creating a dictionary and ways of standardization of the German vocabulary on death certificates were the tasks at hand.

For Morbidity Coding a dictionary of 74,000 entries already existed with normalized permutations of each entry. This dictionary was included in Iris for preliminary tests.

Additionally, Iris offers the possibility to standardize entries through regular expressions (RegEx), a widely used computer science application for finding and matching of defined strings.

As RegEx are a powerful tool to standardize entries we decided to use them sparingly to avoid to much mistakes in the first test.

First results

In the first tests we used the morbidity dictionary only and tested single line certificates to avoid interference of to many factors in one test run. Only very few standardization steps were applied for the special German characters like ä, ö and ü. For the test we used about 250,000 one line certificates, some of them with two or more diagnosis on a single line. The data was provided through local cancer registries in paper format and was manually typed in.

The results were discouraging with only about 30% of one line matches.

Analysis of about 80,000 of these certificates was conducted in order to create new entries for the dictionary and to generate RegEx for the standardization tables of Iris.

MainKey	Rank	FilterIn	FilterOut	Ac	Li	P	DateIn	UserIn	DateOut
300DeleteDoubt	0001	\bverdacht auf					01.02.2006	Weber	
200DeleteDoubt	0002	verdacht\b					01.02.2006	Weber	
200DeleteDoubt	0005	\bv[a-z]*[.]s?a[a-z]*[.]?					01.02.2006	Weber	
200DeleteDoubt	0010	\bmoeglich[a-z]*[.]?					01.02.2006	Weber	
200DeleteDoubt	0015	\bmoegl[a-z]*\b[.]?					01.03.2006	Weber	
300Short	0010	\bobst[a-z]*[.]?	obstruktiv				01.02.2006	Weber	
300Short	0020	\bmalign(nes ner nem heln)\b[.]?	maligne				01.02.2006	Weber	
300Short	0030	\babd(om)?\b[.]?	abdominal				01.02.2006	Weber	
300Short	0040	\babsc(ed)?\b[.]?	abscedierend				01.02.2006	Weber	
300Short	0051	induc([a-z]*)\b[.]?	induciert				31.03.2006	Weber	
300Short	0060	Neub[a-z]*\b[.]?	neubildung				01.02.2006	Weber	
300Short	0070	\bboes[a-z]*\b[.]?	boesartig				01.02.2006	Weber	
300Short	0080	\balcohol(isch ischer ische isches)?\b[.]?	alcohol				01.02.2006	Weber	
300Short	0090	\bAlc[a-z]*[.]?ab[a-z]*\b[.]?	Alcoholabusus				01.02.2006	Weber	
300Short	0100	Abscessaus[.]?[a-z]*\b[.]?	Abscessausraeumung				01.02.2006	Weber	
300Short	0110	\bAlc[.]?[a-z]*\b[.]?[a-z]*\b[.]?	alcoholabhaengigkeitsyndrom				01.02.2006	Weber	
300Short	0120	\bAlc[.]?[a-z]*\b[.]?	alcoholtoxisch				01.02.2006	Weber	
300Short	0121	\balcoholtox[a-z]*\b[.]?	alcoholtoxisch				31.03.2006	Weber	
300Short	0130	\bAlc[.]?[a-z]*\b[.]?	alcoholinduciert				01.02.2006	Weber	
300Short	0140	\bAlc[.]?[a-z]*\b[.]?	alcoholabusus				01.02.2006	Weber	
300Short	0150	\bAlc[.]?[a-z]*\b[.]?	alcoholintoxication				01.02.2006	Weber	
300Short	0160	\bAlc[a-z]*[.]?ab[a-z]*\b[.]?	alcoholabhaengigkeit				01.02.2006	Weber	
300Short	0170	\bAlc[a-z]*[.]?mi[a-z]*\b[.]?	alcoholmissbrauch				01.02.2006	Weber	
300Short	0180	\bAlc[a-z]*[.]?kr[.]?[a-z]*\b[.]?	alcoholcrancheit				01.02.2006	Weber	
300Short	0190	\baltersbed[a-z]*\b[.]?	altersbedingt				01.02.2006	Weber	
300Short	0200	\baltersschw[a-z]*\b[.]?	altersschwaeche				01.02.2006	Weber	
300Short	0210	\balch[a-z]*\b[.]?	alcheimer				01.02.2006	Weber	
300Short	0220	\bamput[a-z]*\b[.]?	amputation				01.02.2006	Weber	
300Short	0230	\baneu[a-z]*[.]?[a-z]*\b[.]?	aneurysma dissecans				01.02.2006	Weber	
300Short	0240	\bangi[a-z]*[.]?[a-z]*\b[.]?	angina pectoris				01.02.2006	Weber	
300Short	0250	\bangeb[a-z]*\b[.]?	angeboren				01.02.2006	Weber	
300Short	0260	\bAort[a-z]*[.]?cl[.]?[a-z]*\b[.]?	aortenclappenstenose				01.02.2006	Weber	
300Short	0270	\baort(en)?cl[.]?[a-z]*\b[.]?	aortenclappen\$2				01.02.2006	Weber	
300Short	0280	\baortoc[a-z]*\b[.]?	aortocoronar				01.02.2006	Weber	

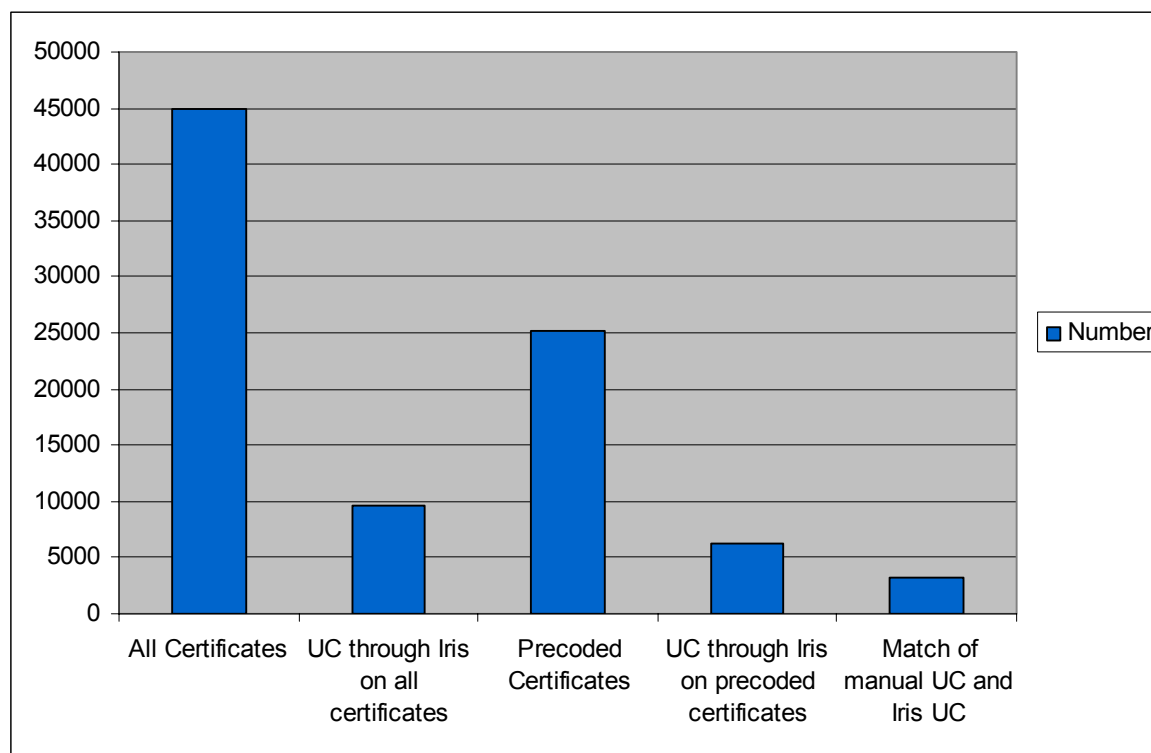
Screenshot 1: Standardization Table in Iris for the German language

Progress in testing

Over the following month various new test were conducted and the rate of automatic recognition of one line certificates rose to 70% over a short period of time. To achieve this rise we added about 200 terms to the existing dictionary and applied about 1,800 steps of standardization.

Still, this resulted in an automated coding of certificates (with an average of 2,3 lines) and assignment of the underlying cause (UC) of only 24%.

Including the results we focused on a test deck of 44,000 certificates with all lines of the certificates that have been filled in. 25,207 of these certificates were manually pre-coded.



Graph 1 Number of Certificates coded through Iris

25% of the manually pre-coded certificates could be coded automatically through Iris. Of these about 53% of the automated UC was identical to the manually assigned UC. This is only a little lower as previous studies on coding variation amongst the 16 counties show.

Analysis

The next step was now to analyze the certificates where Iris-coding did not agree with the manual UC. Furthermore, the certificates that could not be coded automatically through Iris had to be checked.

Therefore we analyzed almost 1,000 certificates so far. We found that about 50% of the assigned manual UC were wrong. This has to be put into perspective as for the test we only used the lines of part 1 and 2 of the certificate and left out the additional information on the certificate, like the epicrisis given through the physician or the "category of accident" – a separate field on the German certificate. Therefore IRIS-coding might have been wrong as well as we did not consider this additional information.

For about 48% of the certificates we did not have a respective entry in the dictionary for at least one word. New texts for the dictionary resulting from the previous work had not been included in this test as they have to be validated according to internal regulations before being entered into the dictionary. We

expect though to have about 6000 new terms ready for the next test and with them lower the percentage significantly.

About 32% of the certificates contained at least one text that was spelled wrong and even 44% of the certificates contained at least one abbreviation. Only a low percentage of these could potentially be corrected through standardization but most of them will have to be corrected manually before automated coding.

Another 4,3% of the certificates could not be coded correctly because they contained external cause information. Here the problem to some extent lies within ACME as it is not giving two codes for these cases: The external cause code and the injury code itself. This problem was discussed in the last Iris Meeting in Alexandria and will be taken care of in one of the next ACME versions.

Problems	Number of certificates with this problem	Total Number of certificates analyzed	%
Manually assigned UC wrong	494	943	52%
One or more expression on certificate not in dictionary	455	953	48%
One or more expression on certificate is spelled wrong	305	953	32%
One or more expression on certificate is abbreviated	420	953	44%

Table 1 Results from analysis of approximately 1,000 certificates

Discussion

The use of Iris with a complicated language like German holds a lot of challenges to be solved. Even with a pre-existing dictionary for morbidity we could not get around intensive work on language standardization and adaptation of the dictionary. Still, as the rise in recognition of single line entries shows, results can be improved a lot with relatively few steps in the first place. Later of course, as standardization and dictionary are already elaborated and efficient a rise in automated assigning of UC can only be achieved in slower steps.

A key role in the level of automatically assigned UC is the way the texts are entered in Iris in the first place. As we did estimate the tests proved that a minimum of 44% of the German death certificates cannot be recognized correctly because of abbreviations and spelling mistakes. This can only be absorbed up to a certain degree through standardization with the risk of misinterpretation of very short abbreviations.

Problem areas of ACME (external causes, maternal mortality etc.) that can not be handled automatically will be passed on to Iris and reject a certain amount of cases.

Even if we can reach only 30% of automated assigning of UC for the first test year in Germany this will still represent 180,000 Certificates per year. With this perspective more time could be spend on tricky cases and standardization of mortality coding for international statistics will be taken a step forward.

With Iris in use in test-areas in Germany starting next year and the ongoing tests in DIMDI we hopefully will obtain further important information on missing terms for the dictionary and on new entries to the standardization tables of Iris.

Dr. Stefanie Weber

WHO-FIC Collaborating Centre for the German Language

German Institute for Medical Documentation and Information (DIMDI)

Waisenhausgasse 36-38A

50676 Köln

Germany

Email: stefanie.weber@dimdi.de

Phone: +49 221 4724 485

Fax: +49 221 4724 444

<http://www.dimdi.de>