SCENARIO TESTING FOR QUALITY LABELING OF ELECTRONIC MEDICAL RECORDS FOR GENERAL PRACTITIONERS

R. Buyl*, M. Nyssen* and A. De Smedt*

* Vrije Universiteit Brussel, Faculty of Medicine and Pharmacy
Department of Biostatistics and Medical Informatics (BISI), Brussels, Belgium

rbuyl@vub.ac.be

Abstract: *Objective:* The main objective of this work was to develop a flexible and widely accepted evaluation method for the quality of electronic medical record (EMR) systems, in the context the long term policy of improved electronic record keeping by the medical professions in Belgium. Efforts lead by the Federal Ministry of Health, implemented via several committees, working groups and teams.
After two years of successful application, we present the methodology, analyse the results and draw conclusions for future developments.
***Method*: The development of a scenario-based test sequence based on a list of requirements or criteria, pertaining professional functionality, user-friendliness and implementation robustness. The new test method was devised and applied during two consecutive years.**
***Results*: We clearly see an increase in the average test-scores in all the categories of the criteria, comparing the two consecutive test sessions, except for the category "data extraction". Moreover, the marked decrease in "spread", shows that the surviving EMR systems become more homogeneous.**
***Discussion*: After two years of in the field practice, we can conclude that we have developed an objective, efficient and easy to use test procedure, well linked to "everyday practice" and accepted by the healthcare professionals.**

Introduction

As recently confirmed by CEN's special focus group on E-Health [17], the electronic medical record (EMR) is considered to be one of the cornerstones of Ehealth. EMRs contribute to the quality of healthcare, they help general practitioners (GPs) in decision making, drug-prescription, but also in exchanging medical data with laboratories, hospitals and other GPs. For example in Denmark about 70% of all medical documents (prescriptions, lab results,…) are exchanged electronically [9].

Similar to medical devices, that require testing and approval by the authorities before being used in the healthcare, the EMR sector will benefit from quality labels, as requirements for the medical software systems [2,4, 6,13].

Therefore in many European countries certification and labeling procedures were started in the 1990's and are still being updated. In the Netherlands the "Referentie model 95" [15] and "MEDEUR" [7] were developed, NHS in the UK made the RFA (requirements for accreditation) [14]. Ireland followed that example and the GPIT (General Practice Information Technology) [5] group completed their new labeling procedure based on NHS's RFA in the around the year 2000. In Denmark too, a project was launched, coordinated by MedCom (the Danish Centre for Health Telematics), for the testing and the accreditation of IT systems, including software packages for GPs [9].

During this period, also in Belgium an accreditation procedure was decided upon, to improve the quality of medical records and thus, healthcare. By giving GPs financial incentives to use "user-friendly" and "quality labelled" packages, the Ministry hopes to achieve this primary goal and to obtain interoperability as an essential by-product.

Materials and Methods

The Belgian Model

Phase 1: Preparatory work (1998-2000)

Under the coordination of the Belgian "Ministry of Social Affairs, Health and Environment" (Federale Overheids Dienst) the "Telematics Commission" was created [1], regrouping several committees. These working groups operate in a transparent way and reach decisions via consensus whenever possible. A huge benevolent effort led to an exhaustive list of some 300 criteria concerning the Electronic Medical Record in Belgium. These 300 criteria were subdivided into 10 major categories as depicted in table 1. In order to guide the vendors in their implementation priorities towards meeting the requirements in a reasonable pace - taking technological border conditions into account - it was decided that every year the criteria would be reviewed and qualified from "absolutely essential" to "desirable in the future". A classification we also find in other countries like Ireland (GPIT) [5]. For the first two test

ISSN: 1727-1983

sessions this resulted in about 100 criteria that were "absolutely essential".

Table 1: Description of the categories.

Cat.	Description
1	General criteria
2	Administrative data
3	Medical data
4	Data structure*
5	Decision support
6	Medical record support
7	Coordination, connectivity, communication
8	Data extraction
9	Medical legal aspects*
10	Overall support and continuity

* Only in the 2003 session

Main developments that had an important impact on the gestation of the criteria:

- An investigation concerning the effective needs and concerns related to the EMR, by general physicians and specialists.
- The results of a working group, designing and proposing the "structure of the EMR". The chosen structure is "patient-centred". It proposes a three-level hierarchy for the EMR, based on the following fundamental concepts: "Health care element", "Health approach" and "Service(s)". The concepts of "(Sub)Contact", "Contact" and "Health agent" define the healthcare context, in which the services are provided [3].
- The advent of XML as a widely accepted data interchange tool and the KMEHR-bis (Kind Messages for Electronic Health Records) XML application, describing health related messages in detail. This development will enable the exchange of structured clinical information [12].

Phase 2: Development of a test suite

Once the list of criteria was available, the next step was the implementation of an objective testing procedure, which would be acceptable by all concerned parties. Therefore, a method which we call "evidence-based testing" was elaborated, relying on scenario-oriented test-suites, built via open fora. The test suites take into account the experience of day-to-day practice of physicians and the list of "essential criteria" requirements. These scenario-based test sequences were elaborated under the coordination of a small independent group of experts through an open forum. Via a top-down design, we started with the most common medical cases and ended up with a set of detailed, well documented test scenarios.

The final test suite was made public, thus avoiding complications resulting from leaks of "inside information". Specific data, concerning names of patients and medications or specific medical parameters were kept confidential. They were made public during the effective testing sessions, to prohibit vendors from completely pre-programming all elements of a scenario session. [8, 10, 13, 16]

The scenarios have following desirable characteristics:

- conforming to the day-to-day practice and emphasising often encountered actions
- relying on real-life patient data
- well identified actions, imposed data entry instructions and strict sequential order of operations
- simple but unavoidable controls
- consensus evaluation of the "true" / "false" type

The use of scenario-testing has another advantage: multiple mapping from an individual test to different elementary criteria, allows making fewer tests in order to evaluate more criteria. This works well, when detailed evaluation forms, pinpointing the observations made in case of failure are carefully filled in by the observers. The report of the test evaluators must allow to relate the failure to one or more unmet criteria. As shown in Table 2, this results in a "testing productivity gain" of about 40%, compared to "a single test per criterion" set-up.

Table 2: category split-up of the 2003 scenario test suite

Item category	**# tests**	**# criteria**
1. Admin and medical data + lab results	6	12
2. Simple registration	7	14
3. Structured registration	16	20
4. Vaccinations	6	7
5. Planning	4	4
6. Set-up and help	3	3
7. Documents, reports	6	11
8. Selections	3	2
9. Back-up and export	4	4
TOTAL	55	77

About a month before the actual test sessions, two documents are published on the Ministry's website:

- the patient records which should be pre-uploaded before the test session
- the test scenario's with pertinent data masked (such as patient names, dates, specific numerical data, medications)

This allows the vendors not only to introduce essential data records at ease, but also to test their packages before the actual official test sessions. These patient records consist of 18 fictitious patients, preloaded into the EMR system and 6 datasets related to lab results/data, half of which are from "known" (belonging to the 18 pre-defined patients), half from "new" patients, to be introduced during the first part of the test session.

ISSN: 1727-1983

During the actual test sessions, these scenarios are completed with pertinent data. Then each package is demonstrated by two experienced users, who are totally independent of the vendor. They can make use of the vendors help desk if needed. To guide the users through the test session a document with scenario instructions is produced. The evaluation is performed by two jury members and one observer, who work in a consensus mode. One of the jury members receives a document with the detailed scenarios, completed with the expected outcomes of all the tasks performed by the users. This document also serves as a scoring booklet on which the jury writes the result of each test, answering with a "yes" or a "no". If a negative outcome is produced, it should documented so that it becomes clear which of the criteria that were tested is not met.

At the beginning of the session, starting time of the test session and the names of all participating persons are recorded carefully on the first page of the booklet. Step by step, the scenario instructions are then read by one of the jury members, executed by the "users" and the results are evaluated and annotated on the score pages by the jury members and the observer. When no consensus for a certain question is reached, this should be also well documented in the scoring booklet.

Example: A typical scenario instruction

Import the data records, available on a disk containing lab results. Also a discharge letter in Kmehr-bis (XML) format has to be read into the system.

T-02.6 Show the complete patient record including all new elements (lab-results, discharge letters,…)

In this example, we test the capabilities of the software package to import data (both administrative and medical data records) from different laboratory sources.

All the data imported from these different sources are temporarily stored in an electronic drop box. At this time it is still possible to view these documents without opening a patient record. The next step tests the package's capability of linking the documents that are now stored in the drop box, to an individual patient.
In our example we now ask the program to show the complete patient record, which already contained some elements before the introduction of these new ones. With this instruction we test the criteria 100, 101, 102, 103 and 104 of the 2003 requirements. To get a positive result the package has to produce the completed dataset, it will then meet all the criteria. If one or more partial data records are missing, a negative result will be noted, extensively documented so that it becomes obvious which of the criteria were not met.

Note also the numbering scheme of the sub-test, identical on all documents, in Dutch and French (here: 6th subtest of the 2th group of tests)

At the end of the test, the produced documents (paper- and electronic) are collected, added to the evaluation booklet, the end time is recorded and everybody involved signs the scoring booklet.

The complete scenario will be built as a sequence of more than 50 instructions such as the one (***T-02.6)***, given as an example above.

Individual instructions are grouped in test categories (here T-02), as reflected in the name of every "atomic" (or undividable) subtest. The complete scenario's are publicly available in Dutch and French, via the Ministry's "homologation" page [18] or via the authors.

Next to the scenario, the labeling procedure comprises a series of contractual and technical constraints for the vendors (guaranteeing support and documentation concerning internal data formats).

Results

A first series of tests and evaluations were performed in 2002, followed by a second series in November 2003. Each time 20 packages were evaluated (not identical sets) in a two day time span. Evaluators filled in the score-forms as the scenarios were produced step by step, allowing to judge the presence or absence of crucial features as required by the «label» criteria. Test duration statistics can be found in Table 3. The longer duration in 2003 stemmed from the larger number of sub-tests (37 in 2002, 55 in 2003) and the complex criteria related to the EMR's structure. No significant relationship was found between the duration of the test and the final outcome (getting the quality label).

Table 3: test durations in 2002 and 2003

	2002	***2003***
Average test duration (min)	95	115
Std. Dev. (min)	25	40
Min-Max. (min)	60-150	70-210

Table 4 and Figure 1 show an overview of the scores obtained by the software packages in 2002 and 2003. The results are presented as an average score (number of test criteria passed) of all packages within every category. Categories 4 (Data structure) and 9 (Medical legal aspects) were not yet included in the 2002 test session. We noticed improving results in 2003 (between 5.83% and 26.64%) in all categories. These differences are significant ($p<0.05$) for the categories 3, 6, 7 and 10. Category 5 (decision support) showed almost the same result in both years. Only category 8 (Data Extraction) scored better in 2002 (85% vs. 75%).

ISSN: 1727-1983

Table 4: Overall results

Test category	Year	result (%)	SD (%)
1. Overall criteria	2002	85.00	23.51
	2003	92.72	12.74
2. Administrative data	2002	90.00	30.78
	2003	95.83	7.08
3. Medical data†	2002	82.31	28.90
	2003	96.42	5.48
5. Decision Support	2002	95.00	22.36
	2003	90.00	30.78
6. Medical record support †	2002	71.25	27.24
	2003	97.89	4.65
7. Coordination, Connectivity, communication.†	2002	70.00	47.02
	2003	92.50	12.06
8. Data extraction	2002	85.00	36.63
	2003	75.00	35.66
10. Overall support and continuity †	2002	80.00	41.04
	2003	99.69	1.40
total score	**2002**	**81.72**	**21.40**
	2003	**91.32**	**10.01**

(† significant with $p \leq 0.05$)

A vendor independent jury then decides on awarding or refusing of the quality label, based on a large overview table, showing the performance of all the packages for all the tested criteria.
After notification of the results, vendors have a chance to make an appeal on the basis of well documented motivations. A few weeks after the test sessions the final results are made public. In 2002 and 2003, 17 of the 20 packages were accorded a quality label (in 2003, 3 packages differed from the 2002 set).

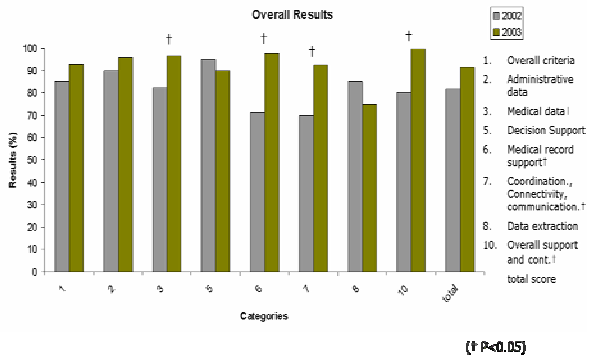

Figure 1: Overall results

Discussion

In Europe, several initiatives are taken to improve and certify the quality of Electronic Medical Record Systems. But we notice that exchangeability of data and interoperability between laboratory-, hospital-, and GP-systems is an area in which still a lot of progress has to be made. This is clearly illustrated by the different types of standards for medical data exchange in the different countries, going from general EDIFACT messages to XML schemes [1, 7]. Nevertheless we are convinced that a crucial and controversial step was taken in the evaluation of the EMR systems in Belgium. Thanks to the collaboration of all involved parties, from the government to the end users and via the software producers, a meaningful momentum towards improved quality has been initiated.

The test results of the first two years are satisfying. We notice good overall results for both years (81.72% for 2002 and 91.32% for 2003). We can see an increase from 2002 to 2003 in almost every category, except for category 8 (Data extraction). This was due to the fact that more attention was paid to this item and most of the criteria tested in this category were not yet obligatory in 2002. Note also the much larger standard deviations in 2002 which indicate the higher heterogeneity of the software packages the first year of testing. We clearly observe that the quality level of the different packages equalised in 2003.

The most important features that were imposed via the label criteria are:

- intrinsic quality of the record and its ergonomics
- the structure of the EMR and its effect on medical recording practice
- interoperability or at least exchangeability of records (also between different vendors)

Conclusions

We can conclude after the first two years of "in the field" application of the "evidence-based scenario testing" method, that we established a procedure that allows efficient and fast evaluation, leaving little or no room for discussion. Moreover we are convinced that the creation of this test suite will contribute in the future development of EMR systems. Standardisation and exchangeability of data will be key issues, that need close attention and steering. In Belgium, a positive momentum was created, involving all interested parties and leading to improved electronic medical records, via a pathway that was designed with the active contribution of all parties involved, including the people actually using the systems. We have created a quality labeling testsuite that will be adaptable to test software systems for dentists, physiotherapists and other health workers such as nurses and emergency personnel. We are convinced that this will lead to a substantial improvement in the interoperability between the different areas of our healthcare system.

ISSN: 1727-1983

Acknowledgements

The authors are grateful to all contributors to the testscenario's, especially the members of the Flemish First Line Forum, the Medical Discussion Forum and to the collaborators of the Ministry, especially Dr. Jean-Paul Dercq, who initiated the ministerial involvement and Dr. Marc Bangels, responsible for the governmental E-health cell, and their teams.

References

[1] DE CLERCQ, P. PIETTE ET AL., *Structure of the Electronic Patient Record*, report of the EMDMI's Structure Working Group, October 21st 2003, Retrieved September 2004, from http://www.health.fgov.be/telematics/label/mh/index.html

[2] DEFTEREOS S. , LAMBRINOUDAKIS C., ET AL., *A Java-based Electronic Healthcare Record software for beta-Thalassaemia,* Retrieved September 2004, http://www.jmir.org/2001/4/e33/index.htm

[3] DIMITRIOS G. KATEHAKIS, STELIOS SFAKIANAKIS ET AL., *An infrastructure for integrated Electronic Health Record services: The Role of XML (Extentible MarkupLanguage)* Retrieved September 2004, http://www.jmir.org/2001/1/e7/index.htm

[4] FORSSTRÖM J (1997), *Why certification of medical software would be useful? ,* Int. J. of Med. Inf, 47, 143-152

[5] General Practice Management System Accreditation, Presentation by Maria O'Brien, GPIT Ireland, Retrieved January 2005 from http://www.gpit.ie/presentations/Accreditation%20for%20General%20Practice.pdf

[6] HAUX R. AND KULIKOWSKI C. (2003), *Yearbook of Medical Informatics 2003 "Quality of Health Care: The Role of Informatics" IMIA*, Schattauer Stuttgart, Eds. , ISBN 3-7945-2263-X, ISSN 0943-4747

[7] Inventarisatie gebruik MEDEUR in de Huisarts Informatie Systemen, Vlug A.E, Retrieved January 2005 from http://www2.eur.nl/fgg/mi/MIDownloads/Documents/Inventarisatie%20gebruik%20Medeur%20in%20HIS.pdf

[8] JOHANNES RYSER AND MARTIN GLINZ, *A Scenario-Based Approach to Validating and Testing Software Systems Using Statecharts*, Dept. Coputer Science Univ. Zurich, Switzerland. (12th International Conference on Software and Systems Engineering and their Applications, ICSSEA'99 Proceedings, CNAM, Paris, France.

[9] JOHANSEN I, HENRIKSEN G, ET AL., (2003), *Quality assurance and certification of Health IT-systems communicating data in primary and secondary health sector*, Presentation on the MIE2003 conference in St. Malo

[10] KANER C., FALK J. ET AL. (1999),*Testing Computer Software*, 2nd Ed., Wiley, ISBN 0-471-35846-0

[11] KIM M.I., JOHNSON K.B, (2002) *Personal health records: evaluation of functionality and utility*, J. Am. Med. Inform. Assoc. 9(2):171-180.

[12] Kind Messages for Electronic Healthcare Records Belgian Implementation Standard, Retrieved September 2004, http://www.health.fgov.be/telematics/kmehr/

[13] NIINIMAKI J., FORSSTROM J. (1997), *Approaches for certification of electronic prescription software*, Int. J. of Med. Inf, 47, 175-182

[14] Primary Care Computer Systems Requirements for Accreditation, Retrieved January 2005 from http://www.nhsia.nhs.uk/sat/specification/pages/pdf/general.pdf

[15] Revisie van het WCIA-HIS-Referentiemodel 95, Retrieved January 2005 from www.mi.unimaas.nl/events/WCIArapport/wciarapport.pdf

[16] Third Annual Workshop on the Teaching of Software Testing (WTST 3) February 6 - 8, 2004 Melbourne, Florida, Retrieved September 2004, http://www.testingeducation.org/conference/wtst_page_2004.php

[17] Current and future standardization issues in the e-Health domain: Achieving interoperability, Report from CEN/ISS e-Health Standardization Focus Group (March 2005), Retrieved April 2005, http://www.centc251.org/ehealthfocusgroup.htm

[18] Belgian Ministry of Health "Homologation of GP packages" web pages , Retrieved March 2005, http://www.health.fgov.be/telematics/label/mh/index.html

ISSN: 1727-1983