Terminological systems and prognostic models as instruments for quality assessment in intensive care
Raiez, F.

Citation for published version (APA):
Raiez, F. (2010). Terminological systems and prognostic models as instruments for quality assessment in intensive care

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
General Discussion
9.1 Introduction

Proper performance measurement is an essential aspect of quality assessment. To this end the quality of performance data and performance indicators are of great importance. The quality of performance data can among others be optimized by standardizing data collection and data processing by means of coded data using a reference terminology. The performance of indicators may be improved by using prognostic models to account for case-mix differences, making performance indicators comparable across providers or institutions. This thesis elaborates on these topics and addresses the use and application of terminological systems and the use and application of prognostic models as instruments to facilitate quality assessment in the intensive care (IC). Therefore three research questions were formulated that were addressed in Chapter 2 to Chapter 8:

1. Can a terminological system be used to standardize diagnosis coding in the intensive care and what characterizes its development, presentation and maintenance?

2. How do prognostic models perform in the Dutch ICUs and does the choice of a prognostic model and its customization affect the results of quality assessment?

3. What is the impact of using SNOMED CT to crossmap two intensive care-specific classification systems on quality assessment?

This final chapter provides a summary and discussion of the work presented in the former chapters. For each research question the principal findings are provided, the strengths and weaknesses of the study results are discussed in relation to the results of other studies in the field, and topics for future research are identified throughout the chapter. The chapter concludes with the implications of the study for clinical practice.

9.2 Standardized diagnosis coding by means of terminological systems

9.2.1 Summary of Principal Findings

The first research question was divided in four sub questions which were described in Chapters 2 to Chapter 5.

Can a terminological system be used to standardize diagnosis coding in the intensive care?

To evaluate the practicability of terminological systems for recording clinical encounters, the study described in Chapter 2 focused on whether clinicians in day-to-day care practice
can adequately capture reasons for IC admission by using a terminological system. A locally developed compositional terminological system was integrated in the Patient Data Management System (PDMS) for structured registration of reasons for ICU admission. The correctness and specificity of the terminological-system-based reasons for ICU admission were comparable to the correctness and specificity of the free-text reasons for ICU admission. Consequently, in view of the many benefits of the terminological-system-based registration, the study concluded that free-text recording of ICU reasons for admission should preferably be replaced by terminological-system-based recording to standardize diagnosis coding in the intensive care.

What characterizes the development of a terminological system?
In spite of the positive results, a full implementation of the local terminological system as described in Chapter 2 was not realized, as its completeness and integrity could not be guaranteed, and its maintenance was found to be labor-intensive and time-consuming. Instead, an interface terminology based on SNOMED CT was considered to standardize diagnosis coding in the intensive care. Using a reference terminology such as SNOMED CT rather than a home grown interface terminology is considered to improve semantic interoperability, i.e. the ability to exchange data without changing the meaning, and thereby supports data-reuse (1;2). The development of this interface terminology was addressed in Chapter 3. This study defined a generic approach for developing a domain-specific interface terminology based on SNOMED CT. This process was regarded as consisting of six sequential phases: domain analysis, mapping from the domain concepts to SNOMED CT concepts, creating the SNOMED CT subset guided by the mapping, extending the subset with non-covered concepts, constraining the subset by removing irrelevant content, and deploying the subset in a terminology server. This approach was applied to develop an intensive-care-specific interface terminology on SNOMED CT.

What characterizes the presentation of a terminological system?
The IC specific interface terminology from Chapter 3 was evaluated in clinical practice for data entry using the terminology client as described in Chapter 2. The results were described in Chapter 4. The ease of use of a terminological system is influenced by many factors, such as its complexity and language. Overall, the study revealed that the terminology client fell short of the users’ expectations. A high number of distinct usability problems were identified which resulted in ineffective and inefficient registration of reasons for admissions. A majority of the encountered usability problems was related to the terminology client. However, the problems related to the interface terminology based on SNOMED CT were more severe. The study concluded that to encourage the clinicians to code the clinical encounters systematically using a large compositional interface
terminology, the interface terminology should be presented to the users in such a way that it is possible to capture clinical entries with minimal effort. The study suggested adjusted subsets, enhanced terminology clients with advanced search and navigation mechanisms and term selection, and technologies such as natural language processing (NLP) to facilitate data entry in clinical practice.

What characterizes the maintenance of a terminological system?

As described above, the local terminological system as described in Chapter 2 was not implemented because its completeness and integrity could not be guaranteed and local maintenance was found to be labor-intensive and time-consuming. This emphasizes the importance of a well organized maintenance process. Chapter 5 provided a framework that summarized the required features of a maintenance process of a terminological system for sustaining the accuracy of the content of the terminological system. The framework was used to explore the current practice of terminological system maintenance by a survey among organizations that maintain a terminological system. The survey showed that maintenance of larger terminological systems fulfills most of the criteria mentioned in the framework whereas maintenance of smaller terminological systems fails more criteria, probably as this incurs costs and overhead that are too high and/or unnecessary. The survey also revealed that the framework is applicable to all kinds of medical terminological systems and provides maintainers of medical terminological systems with the principles needed to have a well-organized maintenance process.

9.2.2 Strengths and weaknesses of the study and related research

Terminological systems aim to facilitate the recording and retrieval of clinical information. It has been argued that without the use of a standard terminological system, the full benefits of a PDMS, i.e. effective use and communication of health care information within and across organizations, are unlikely to be realized (3;4). Therefore, in recent years, the use of reference terminologies and in particular SNOMED CT has received growing attention. A large part of the research in this area is concerned with an evaluation of the content coverage of SNOMED CT to adequately represent clinical encounters, or with the use of SNOMED CT for automatic coding of clinical encounters involving some kind of natural language processing technique (5). Few studies have investigated the use of SNOMED CT in clinical practice for data entry. The impact on care assessment of terminological-system-based registration of clinical encounters is as yet underexposed.

A strength of our research is that we considered the entire process of terminological-system-based registration, with special focus on SNOMED CT, i.e. utility, development of an interface terminology, its presentation to the users for data entry, and the terminological system’s maintenance. The methodologies and lessons learned provide a valuable
foundation for the further application of terminological systems in clinical practice. However, there are some limitations of our current work that we should address.

One of the advantages of using a subset of SNOMED CT in the interface terminology was to isolate the users from the complexity of SNOMED CT and to extend the subset with user-required content which was not included in SNOMED CT, e.g. local synonyms. Ideally, the subset should be as compact as possible while at the same time including the level of granularity needed in the clinical domain of interest (6,7). Depending on subsetting technique, as defined in this thesis, the size of the subset could vary between 5,500 and more than 80,000 concepts. When fully expanding the subset, more than 150,000 concepts had to be added as attribute values to further specify the disorder and procedure concepts. The domain of intensive care is rather complex and broad, involving a diversity of clinical problems. So, for the interface terminology to facilitate sharing and aggregating this kind of data for different purposes, it should enable a detailed registration of widely diverging clinical problems and their attributes. Therefore, we chose to use the fully expanded subset.

A previous study showed that a subset of about 2,700 concepts from SNOMED CT was sufficient to (retrospectively) code 96% of clinical notes for patients admitted to an Australian intensive care over a period of 5 years (7). Although this study was performed in another setting with a different basic assumption, it shows that the size of our interface terminology was rather large. This was also confirmed in the usability study in which the extensiveness and complexity of the SNOMED CT subset appeared to be an impediment for data entry. Compared to the restricted home-grown terminological system as described in Chapter 2, the users had more difficulties to code reasons for admissions using the SNOMED CT subset deployed in the same terminology client. Hence, the interface terminology in its current form is too large and too complex to be usable for data entry in clinical practice.

Probably the most promising way to standardize diagnosis coding in the intensive care using SNOMED CT is automated coding by NLP algorithms to encode free-text input (8-13). In general, the use of NLP in clinical practice yields the promise to shield the users from the complexity of the underlying reference terminologies such as SNOMED CT and to ease the data entry process. The use of NLP could overcome most of the problems related to the terminology client as identified in this thesis. However, there are also limitations to the use of NLP. First of all, the interface clients used for NLP algorithms would also require enhanced components to automatically code the clinical information in free text and to suggest relevant SNOMED CT concepts to the users for confirmation. Furthermore, to facilitate a detailed registration of clinical problems based on a compositional terminological system, these interface clients would also need components to stimulate clinicians to further specify insufficiently detailed clinical entries through post-coordination. Besides, the tools used in the available client interfaces for NLP are usually
domain and language specific. As no Dutch translation of SNOMED CT exists, building these tools to process Dutch terms might be too costly.

A major advantage of using a subset of a reference terminology such as SNOMED CT instead of a home-grown terminology is its maintenance. This enables for instance to take full advantage of SNOMED CT’s content and editorial board efforts. However, a new version of SNOMED CT is published every six months and may introduce (large) changes in the terminology content (10;12;14). These changes can affect the local (extended) subsets and consequently the consistency of the captured data. Therefore, also the interface terminology needs to be updated on a regular basis and an update mechanism should be in place to organize this process. A strength of the framework for the maintenance process of the medical terminological systems, as provided in this thesis is that it is applicable to a broad range of terminological systems. At the same time, because of its generality, it covers the characteristics that are relevant for specific types of terminological systems e.g. interface terminologies, in a rather limited way. For instance, the changes in the core reference terminology and the consistency of the mappings from the interface terminology to the core reference terminology are not considered.

9.2.3 Future work

The emphasis of this part of the research was on the practicability, development, presentation and maintenance of terminological systems for clinical practice. Following the discussion, future work should build on these results and focus on the further refinement and application of the methods described in this thesis.

The interface terminology based on SNOMED CT in its current form proved to be too cumbersome to be usable for data entry in clinical practice. Therefore, the usefulness of NLP techniques to ease the data entry based on a compositional terminological system in clinical practice should be investigated.

It has been argued that, next to terminology content, correctness and specificity of terminology-based registration in clinical settings also depend on the presentation of the interface terminology to the users (15). Studying the interface terminology and the terminology client together offered the possibility to evaluate what the exact causes of failures were, i.e. terminology content or client. However, in our study we only evaluated one user interface for real-time data entry. In the future, a better approach will be to test the interface terminology in multiple user interfaces, i.e. involving direct data entry versus some kind of NLP. By doing so, it becomes possible to determine the best practice for terminological-system-based data entry in clinical settings.

Finally, to safeguard the utility of interface terminologies for clinical practice, it is necessary that its maintenance is performed in an organized and structured way. Therefore, future work should focus on the refinement of the framework for the maintenance of
medical terminological systems for the different types of terminological systems such as interface terminologies.

9.3 Standardized quality indicators by means of prognostic models

9.3.1 Summary of Principal Findings

*How do prognostic models perform in the Dutch ICUs and does the choice of a prognostic model and its customization affect the results of quality assessment?*

To address the second research question Chapter 6 firstly focused on the influence of the choice of a prognostic model (i.e. The Acute Physiology and Chronic Health Evaluation II (APACHE II) model, Simplified Acute Physiology Score II (SAPS II) model and Mortality Probability Models II (MPMII)) on league tables (i.e. rank-order listings) in which ICUs are ranked by Standardized Mortality Ratios (SMR). It was shown that assessments of institutional performance in case-mix adjusted league tables are sensitive to the model that is used for case-mix adjustment, irrespective of the customization of the models.

With the introduction of the Acute Physiology and Chronic Health Evaluation IV (APACHE IV) model in 2006, it was suggested that the older models such as APACHE II and SAPS II should no longer be used because they become increasingly inaccurate as the time between their development and application increases (16). Besides, the APACHE IV model incorporates a large number of diagnoses which makes it possible to analyze mortality outcomes in specific subgroups. The APACHE IV model makes it also possible to make mortality predictions for patients admitted after Coronary Artery Bypass Graft (CABG) surgery who were excluded in the APACHE II and SAPS II models. Chapter 7 elaborated on this and focused on the evaluation of the prognostic reliability of the APACHE IV model in the Dutch intensive care population and compared the performance of the APACHE IV model to the performance of the currently applied APACHE II and SAPS II models in different sub-populations. Overall, none of the models outperformed the other two models. The overall discrimination and accuracy of the customized APACHE IV model were statistically better and the overall calibration was inferior to those of the customized APACHE II and SAPS II, but these differences were small in perspective of clinical use. So, in the view of prognostic reliability, after recalibration, all three models could equally be used for quality assessment purposes.

9.3.2 Strengths and weaknesses of the study and related research

The use of performance indicators as a basis for performance assessment and the use of prognostic models to obtain case-mix adjusted performance estimates has been the subject in many studies (12,17-23). Research has shown that different prognostic models may
produce different impressions about care performance, especially when prognostic reliability is indefinite (18;24). Reporting performance thus may provide misinformation. This has important implications as payers and consumers of health care are increasingly seeking information on quality of care in performance-based league tables.

The research presented in this thesis applied powerful statistical methods to evaluate the prognostic models for the Dutch ICU population and to reliably judge the influence of the used prognostic model (and its customization) on quality assessment in performance-based league tables. The absence of a gold standard made it however impossible to quantify the validity of the individual prognostic models for quality assessment in performance-based league tables. Instead, the study focused on how strongly different prognostic models (dis-) agree on institutional comparison of ICU.

A shortcoming of our study was that at the time of evaluating the influence of the choice of prognostic models on quality assessment, the APACHE IV model was not yet available and could not be included in the study (25). However, unpublished results based on recent data showed that also the performance-based league tables associated with the APACHE IV model were inconsistent to those associated with APACHE II and SAPS II models. Because of the surplus value of the new APACHE IV model, i.e. large number of diagnoses and prediction of mortality for CABG patients, the APACHE IV model may be chosen as the preferred model for case-mix correction and quality assessment. Yet, the increasing use of electronic interfaces will minimize the difference in data collection burden between the APACHE II and APACHE IV models. The realization of the mapping between the APACHE IV classification and APACHE II classification as summarized in Section 4 will for instance contribute to less registration burden.

Although in view of prognostic performance all three models could be used for quality assessment, the importance of differences in performance based-league tables based on these models should not be underestimated. If rank data are used by ICUs internally as an indication of their performance compared to other ICUs, the disagreements between these models may be of limited importance. External use of performance data to judge quality of care or to steer national health care decisions makes these differences between the models very important. Therefore, any application of league tables must be with great caution and performance information in these league tables should not be used to draw definite conclusions (26-29).

9.3.3 Future work

The study described in Chapter 6 only evaluated whether different prognostic models and customization strategies lead to different rankings of ICUs, but it did not assess which prognostic model and which customization strategy lead to the most accurate ranking. In order to make such an assessment, a simulation study is needed (30). Simulation studies are
computer intensive procedures to assess the performance of statistical methods in relation to a known truth. Such evaluation cannot be achieved with studies of real data alone. To determine which prognostic model leads to the most accurate ranking of ICUs in terms of case-mix adjusted mortality, a simulation study should be conducted with artificial, manipulated patient outcomes. To our knowledge, such a study has not been undertaken at present. A particular point of interest would be the relationship between common measures of predictive performance and ranking accuracy.

9.4 Effect of standardized data on quality assessment

9.4.1 Summary of principal findings
The APACHE IV prognostic model is almost backward compatible with its predecessor, the APACHE II prognostic model, meaning that co-variates for APACHE II prognostic model are also used in the APACHE IV prognostic model. The covariate “reasons for IC admission” forms an exception in that it requires another classification system. Therefore, the primary reasons for IC admissions in the National Intensive Care Evaluation (NICE) registry are currently collected based on both versions of the APACHE classification, resulting in inefficient double registration. A cross map between these successive versions of classification systems was necessary to maintain the continuity of quality assessment in the NICE registry. The study in Chapter 8 showed that SNOMED CT can be used as a mediator to standardize the crossmapping between different versions of classification systems. In this study, the SNOMED CT crossmap was compared with a manual expert-based and a data-driven crossmap. For a majority of the most frequently occurring APACHE IV categories, the three mapping strategies resulted in the same APACHE II categories. In other cases, there was an overlap between the SNOMED CT matches and the matches provided by one the two other strategies. Differences in the crossmap results had however no significant influence on the predicted mortalities. An advantage of using SNOMED CT crossmapping is that once the mapping is generated with the target classification system, it can be re-used to identify crossmaps to all other classification systems that are already aligned to SNOMED CT.

9.4.2 Strengths and weaknesses of the study and related research
A crossmap between (successive versions of) terminological systems and classification systems is necessary to enable reuse of clinical data collected for one purpose for another purpose, to preserve the continuity of health care documentation due to versioning and, to avoid entering data multiple times, and the associated risk of increased errors (31).
For instance, reference terminologies such as SNOMED-CT are aimed to facilitate data entry in the PDMS during the course of patient stay. Reference terminologies are inadequate for serving the purposes for which classification systems, e.g. the International Classification of Diseases (ICD) and APACHE II/IV reasons for admission, are used. The benefits of using SNOMED CT increase if the reference terminology is mapped to standard classification systems. Such mapping enables data re-use for the purpose of generating health information necessary for statistical analysis and reimbursement. Furthermore, the 10th revision of the International Classification of Diseases, ICD-10, for instance, was developed in 1993 to replace ICD-9, which stems from the 1970s. Many changes were made to coding in ICD-10-CM (32). These changes have a direct impact on the health care systems in that the changes produce inconsistencies and discontinuities in the health care statistics (33-35). A crossmap between such classification systems, would assist the continuity and interoperability of health care documentation and avoid double registration. To our knowledge, we are the first to evaluate the use of SNOMED CT as an intermediate to generate such a crossmap between two specific classification systems and to evaluate the influence of this mapping on quality assessment.

A possible drawback of the use of SNOMED CT as an intermediary is that it may be time consuming and labor intensive, as first a mapping needs to be created from the source classification systems to SNOMED CT. In our case, the mappings from APACHE II and APACHE IV classifications to SNOMED CT were manageable. However, for large classification systems such as ICD, the efforts are incalculable, e.g. ICD-9-CM contains about 13,000 diagnosis codes, while ICD-10-CM has more than 68,000, and SNOMED CT more than 310,000 medical concepts. Furthermore, previous studies on mapping a classification to SNOMED CT have shown that the mapping can be influenced by the structure and content characteristics of both systems (36;37). Due to this, a 1-to-1 mapping between SNOMED CT and the classification systems might not be possible. In these cases, an n-to-n mapping is generated which requires an additional decision to select the appropriate target links (38). This makes the automation of the mapping process challenging (Sue Bowman), especially for large classification systems such as ICD. Nevertheless, to preserve continuity of health care data, these mappings are essential (33-35). The replacement of ICD-9 by ICD-10 makes it even more important to further evaluate the utility of SNOMED CT as a cross map between classification systems and future research should focus on this.

9.5 Implications of the study and final remarks

The increasing need for more accountability and transparency in health care results in a greater release of health information. Therefore, reliable performance data and performance
indicators in clinical registries such as NICE are imperative. The research presented in this thesis builds on research in the domains of quality measurement and data standardization. It highlights two important instruments to enable quality assessment in the Dutch ICUs, i.e. terminological systems and prognostic models.

Standardized clinical terminological systems such as SNOMED CT are considered as the backbone of the PDMS in reducing the variability in the way data are captured and encoded. To rule out inefficient registration in the Dutch ICUs, clinical information should preferably be captured on a detailed level in day-to-day practice using a reference terminology such as SNOMED CT. Adoption and implementation of SNOMED CT will facilitate health information networks and interoperable exchange and use of health information between PDMSs and other health care information systems. This study showed that physicians can use a compositional terminological system to record clinical data. However, further research is needed to fully explore the possibilities for implementing SNOMED CT or other reference terminologies in the PDMS in Dutch ICUs.

In the mean time, the registration burden for the NICE participants can be reduced by applying the SNOMED CT crossmap between the APACHE IV and APACHE II classifications. By using this mapping, the continuity of health care assessment based on each APACHE II or APACHE IV prognostic models is preserved and inefficient capture of the related reasons for ICU admission is prevented.

Prognostic models such as APACHE II and APACHE IV often do not adjust for all possible risk factors and different prognostic models provide different impressions about care performance. Therefore, because in today’s health care environment the use of such prognostic models has important implications, application and validation of these prognostic models for case-mix adjustment is complicated. Most professionals agree that quality indicators based on these models should only be concerned as screening tools that prompt further investigation. For instance, the (mis-) use of league tables that do not take adequate account for confounding factors and uncertainty related to the ranks can be seriously damaging for institutions and may mislead the public opinion. The results of this thesis even more underline the importance of realizing that quality indicators based on prognostic models in the NICE registry should not be used for judgment (e.g. accountability) but for improvement (e.g. change initiatives). As long as performance data are not optimized and performance measurements are not standardized, publication of performance indicators may provide misleading information and may have important implications for payers and consumers of health care. Even then, any application of performance indicators must be of great caution and performance information should not be used to draw definite conclusions.

This increasing importance of quality assessment and publication underlines the necessity for standardized performance data and performance indicators. We believe that the research
presented in this thesis provides valuable insight in the deployment of terminological systems for standardized data collection and the use and application of prognostic models as instruments to facilitate quality assessment in the intensive care.

Bibliography