Information for intensive care evolution: methods to assess and improve data quality and data processing

Arts, D.G.T.

Citation for published version (APA):

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: https://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.
Chapter 8

General discussion
The topics of this thesis are quality of data captured in medical registries and quality of data processing for performance measurement. In this thesis we focus on the field of intensive care. During our research we have therefore performed case-studies on the Dutch National Intensive Care Evaluation (NICE) registry, the Diagnoses for Intensive Care Evaluation (DICE) terminological system, and severity-of-illness scoring systems and prognostic model that were specifically developed for intensive care patients. Nevertheless we aim to provide valuable information for medical registries, terminological systems and prognostic models in general.

In this last chapter the main results and the implications of our research will be discussed. Firstly we will consider the causes of errors during data collection and the measures that can be taken in order to avoid them. In addition we will discuss our findings regarding the accuracy of severity-of-illness scores and in-hospital mortality probabilities. Secondly we will discuss the results of our study on methods for the evaluation of the content of medical terminological system. Thirdly we will discuss the external validation of prognostic models and the effect of the size of the external validation sample on the measured performance of the prognostic models. Directions for further research are proposed throughout the chapter. The chapter concludes with some general remarks and some remarks specifically directed at the NICE registry and the DICE terminological system.

**Quality of data collection in medical registries**

*Causes and types of data errors*

To identify the causes of inaccurate or incomplete data in medical registries we have performed a literature study and a case-study (chapter 2). Data errors appeared to occur regardless of the data collection method (manual or automatic) and at every step of the collection process. From our case study it appeared that in case of automatic data collection data errors are mostly systematic, whereas in case of manual data collection errors were mostly random. Random data errors occurred during the recording of the data on the case record forms, due to inaccurate transcription or non-adherence to data definitions. These types of errors might also have occurred during the capturing of data in the Patient Data Management System (PDMS). However, in our case study, when data were collected automatically, we did not analyse the quality of data in the PDMS. Instead, in case of automatic data collection, we considered the data in the PDMS to be the gold standard data.
General discussion

Data quality assurance

To minimize inaccurate and incomplete data in medical registries several measures can be taken. We have made an inventory of procedures that are aimed at quality assurance of data capturing, and placed these procedures in a framework (chapter 2). The process of data quality assurance should include a continuous feedback loop. Within the continuous data quality assurance process the framework can be useful for reviewing procedures and identifying possibilities for improvement of data quality. Identification of possibilities for improvement should then be based on identified causes of data errors. The procedures that are included in the framework were abstracted from literature. For most of the procedures no empirical evidence of their additional value to the improvement of data quality was provided. Pre-and post-measurements of data quality need to be conducted to determine whether implementation of the framework in fact does reduce the percentage of data errors. The positive effect on data quality of one of the procedures in the framework, the training in data definitions and data collection guidelines, was proven in chapter 3. In this study we evaluated the short-term effects of training on data quality. Further studies would be necessary to determine how long these beneficial effects last.

Quality of data in the NICE registry

A number of the quality assurance procedures proposed in the framework in chapter 2 have already been implemented in the NICE registry. For example, a data dictionary exists, containing definitions of all data items that are documented in the registry. In addition, all participants of the NICE registry are obliged to attend a training in collecting the data accurately according to the stated data definitions. Both locally and centrally the data are automatically checked for range and consistency.

Another quality assurance measure included in the framework is the performance of data audits. In order to assess the data quality in the NICE registry and to see if the current quality assurance procedures are sufficient we have performed data audits at nine centres that participate in the NICE registry (chapter 4).

Two centers that were visited collect their data automatically. In these two centers the number of errors in the documented data was relatively high, due to incomplete extraction software or incomplete source data. The fact that the centers in question were able to solve these error causes and have thus been able to increase the quality of their documented data indicates that the site visits and data audits are useful.

Data in the NICE registry are mainly used for severity-of-illness scoring and for estimating the probability of in-hospital death, among others by means of the
APACHE II [1] and the SAPS II [2] models. We have therefore specifically analysed the effect of the identified data errors on these scores and probabilities. For both the scores and the probabilities of in-hospital death very high agreement was found between those based on the registry data and those based on the re-abstracted data. This indicates that the NICE registry data are of good quality in view of the intended use: case-mix correction for quality assessment of intensive care. However, some of the individual variables that are documented in the NICE registry showed relatively large numbers of missing or incorrect values. Examples of these variables were the alveolar-arterial oxygen difference (A-aDO2) and the daily urine output. Whereas the relatively high error rates for these variables did not significantly affect the severity of illness scores and mortality probabilities, they may cause considerable problems when they are used for other purposes such as epidemiological research.

Recommendations can be made in order to further improve data quality in the NICE registry. For example the NICE coordinating center should see more to it that stated guidelines are followed by the data collectors, and that identified incomplete or incorrect data are corrected afterwards. The site visits to centres participating in the NICE registry was performed for the first time. Consequently we could not compare our results to those of previous site visits. To enable comparison of data quality over different time periods and because each year new centres participate in the NICE registry we do advise to perform site visits more regularly, e.g., once every two years.

**Quality of severity-of-illness scoring**

A relatively new scoring system for quantification of severity of illness is the Sequential Organ Failure Assessment (SOFA) scoring system [3]. The developers of the SOFA scoring system aimed to develop a system that defines the degree of organ dysfunction based on a limited number of simple and objective variables. With a total of 12 variables the SOFA score contains less variables than most other severity-of-illness scoring systems, such as APACHE II and SAPS II. The SOFA scoring system contains only one subjective variable, the Glasgow Coma Score (GCS) for quantification of neurological failure. Other scoring systems also include for example reasons for admission and co-morbidities, which are very sensible for inter-rater variability. Furthermore, some of the other scoring systems require that the physician for some variables chooses between the highest and the lowest value in order to select the worst value. A study by Holt et al. [4] has shown that this causes a large amount of scoring errors. The SOFA score only requires per variable either the lowest or the highest values and does not require physicians to choose.

We have assessed the reliability and the accuracy of SOFA scoring by intensive care physicians (Chapter 5). From this study SOFA scoring by physicians
appeared to be reliable. However, the subjective component of the SOFA score, the GCS, showed relatively low reliability and accuracy.

Low inter-rater agreement for GCS scoring was also shown in previous studies [4-7]. The difficulty of GCS scoring is due to the frequent use of intubation, ventilation and sedation in intensive care. Some physicians assess a patient before sedation, whereas others simply document a normal condition. A strict guideline should state how to score GCS in case of sedated patients. A guideline, that says to assess the probable neurological status without sedation, has already been implemented in the NICE registry.

We can conclude that the SOFA scoring system is indeed somewhat simpler than other common scoring models in intensive care. However, it does include some variables, also included in many of the other models, that cause disagreement and inaccuracy. Nevertheless, as long as better ways to assess organ dysfunction are not available, accuracy of severity of illness scoring may be further improved by measures such as extensive training in data definitions and scoring rules and explicit scoring rules.

In addition, some of the causes of errors in SOFA scoring that were identified in our study, e.g., calculation errors, would not have occurred if SOFA scores were calculated automatically based on electronically stored data in a PDMS. On the other hand, using electronic patient records to automatically derive severity of illness scores introduces new problems; higher scores due to a higher sample frequency and measurement errors [8,9], while some variables such as GCS are subjective and therefore not automatically extractable [10]. Another limitation of computerized severity of illness scoring is the often use of free text in the electronic medical record. For example the APACHE II prognostic model is, among others based on the patient's reason for admission to the ICU. The APACHE II prognostic model provides a list of 54 diagnostic categories. Up till now the patient's reason for admission is frequently documented in free text in the patient record. Consequently the reason for admission can not be processed by a computer algorithm in order to automatically select the correct APACHE II diagnostic category. Use of the APACHE II structured list of diagnoses for recording information about daily care practice is also not desirable, because this list does not provide enough detail. The terminology system DICE can be helpful by enabling a standardized and structured registration of patient's reasons for admission to the ICU at a sufficient level of detail. A prerequisite is that the reasons for admission in DICE are mapped to the diagnostic categories that are used in the APACHE II scoring systems. We are currently working on ways to achieve this mapping.
Standardized data by means of medical terminological systems

Methods for evaluation of the content of terminological systems

Next to the possibility of providing mappings to, for example, classification systems that are used in severity of illness scoring, terminological systems may have other advantages. For example, the structured documentation of the patient’s reason for admission also enables aggregation of homogenous patient groups with respect to their reason for admission. This can be useful for stratification and for selection of patient groups for management and research purposes. In any case, to be useful in care practice the content of a terminological system needs to fulfill a number of requirements [11,12]. Of these requirements, we have focussed on the completeness and the correctness of a terminological system’s content (chapter 6).

Three common evaluation methods that focus on the coverage and the correctness of a terminological system’s content were presented and applied in a case study to the TS DICE. From this study it became clear that each method has its strengths and weaknesses. We advised that the methods are used in combination with each other.

Manual review of the terminological system’s content appeared to be labour-intensive and thereby time-consuming. However, it also appeared to be a very valuable evaluation method, considering the relative large amount of errors and omissions uncovered by this method. This has incited us to further explore the possibilities of this evaluation method. We have developed a computer application, called KEBoRT (Knowledge Editorial Board online Reviewing Tool) [13], that enables domain experts to review a terminological system’s content through the internet. Domain experts can give comments if they feel something is missing or incorrect. In addition they can view each other’s comments and indicate whether they agree with it. By agreeing with each other’s comments consensus can be reached upon the changes that need to be made in the terminological system. The reviewing system also provides automatically generated reports of the comments given by the domain experts and thus facilitates the job of the content manager. The first uses of the reviewing system by intensive care physicians for reviewing the content of DICE are very promising.

Another evaluation method, the formal algorithmic evaluation, identified a relatively small number of errors. Nevertheless we believe that further exploring of this evaluation method might increase its sensitivity. If sensitivity of this method can be improved then it has the potential to focus the efforts of human reviewers and thereby decrease their workload. To exploit the possibilities of formal algorithmic evaluation additional research has been and is still being performed at our department [14-16]. In the case study described in chapter six
we analysed three different evaluation methods. We are aware of the fact that other evaluation methods such as lexical matching [17] exist or will be developed in the future. Additional studies should be performed in order to analyse these other evaluation methods.

Our case study has also shown that the terminological system DICE is not yet complete and free of errors. The errors and omissions identified by our case study were corrected or added. However, since the evaluations in the case study were all based on random samples, this indicates that still diagnoses are missing or incorrectly defined or classified. Actions should be undertaken to identify these errors and omissions. A number of intensive care physicians have already started reviewing the current content of DICE by means of KEBoRT. In addition, we expect that a large part of the missing diagnoses will be identified as the system is being used in real practice. We have developed procedures that include the collection of diagnoses that need to be added and the adding of these diagnoses to the DICE content. Within these procedures the previously mentioned reviewing system KEBoRT fulfils a central role. KEBoRT will be used for the reviewing of and reaching consensus upon the proposed changes. Changes in the content of DICE, such as the inclusion of new concepts or additional terms, will only be made if domain experts have reached consensus. Provided that the improvement of the DICE content will be a continuing and well designed process we believe that DICE can be very useful in real practice and valuable in the total infrastructure for assessing the quality of intensive care.

From data to quality indicators prognostic models

External validation of prognostic models and the effect of sample size

Up till now we have discussed studies that aimed at complete, accurate and structured documentation of data. This is of course very important if the data is to be used, e.g., for evaluative research. However, data of good quality does not guarantee that the processing of the data into information is valid. In the context of intensive care we process the data that is documented in the registry into estimates of the in-hospital mortality, based on prognostic models such as the APACHE II model and the SAPS II model. By dividing the actual in-hospital mortality by the estimated mortality we retrieve the Standardized Mortality Ratio (SMR), which provides information regarding the quality of the delivered care. If the SMR is used as a quality indicator than we need to be sure that the estimated mortality is valid for the population in question. Unfortunately, prognostic models do not always perform as well for other populations as for
General discussion

the one on whose data the models were developed. Therefore, prognostic models need to be validated for new populations.

Several of such external validation studies have been published in the field of intensive care [18-22]. The results of these performance measurements that were published in the past 10 years differ with respect to the best performing model and the decision whether to apply a model or not. The results of our study indicated that these differences can at least partly be explained by the varying sizes of the validation datasets that were used in these studies (chapter 7). It appeared that with smaller datasets the measured accuracy and discriminative ability of a model is subject to large variation, which makes the measurements unreliable. Larger validation datasets ensure the reliability of the measured accuracy and discrimination, but have a negative influence on the measured calibration of a prognostic model.

For the validation of the original APACHE II, SAPS II, MPM₀ II and MPM₂₄ II [23] models we had the availability of a sample consisting of 42139 observations. Overall, the SAPS II model and the MPM₂₄ II model appeared to perform best. Calibration of all the models was found to be significantly insufficient. In order to improve the performance of the prognostic models for use in daily practice or in quality assessment studies in the Netherlands we advise to customize the models for the NICE population [24,25].

Concluding remarks

General remarks

In the near future assessment of quality of health care will remain important for patients, health care providers, health insurance companies as well as governments. The collection and processing of data form the basis for quality assessment. We have identified several possibilities for ensuring the quality of the data collection process. Even though the added value of our framework for data quality assurance has not yet been proven we believe that it does provide a good reference for existing and coming registries. One has to realize though that the process of data quality assurance is a continuing process. Participants and coordinators of medical registries should continue to identify data errors and error causes, and act upon these throughout the existence of the registry. This is also true for medical terminological systems. Continuous reviewing and maintenance of a terminological system’s content is a prerequisite for keeping it accurate, complete and up to date. In addition we have paid attention to the validation of prognostic models that are used for processing the data into indicators for the quality of health care. Reliable assessment of quality of care is not ensured by high data quality alone but also by the validity of the data.
General discussion

processing. Our study provided valuable recommendations for future validation studies of prognostic models.

Remarks concerning the NICE registry

Provided that the NICE registry will continue the quality assurance of their data we believe that they provide a valuable data source. The number of participating intensive care units in the NICE registry is still increasing, making the national database more valuable, but also making task of data quality assurance more complex. It is the task of the NICE board and of the coordinating center to ensure that the current level of data quality is maintained or improved. In addition, use of the NICE registry data for quality assessment projects, and also for other types of research should stimulate the participants’ awareness of the need for high quality data. The APACHE II, SAPS II, and MPM II severity-of-illness scoring systems and prognostic models will continue to play an important role in quality assessment of intensive care. The current size of the NICE registry database enables a complex customization of the models, which will be carried out in the near future.

Considering the developments in data quality assurance and prognostic model customization, the intended integration of the DICE terminological system and the continuing increase in the number of participating hospitals the NICE registry provides a valuable infrastructure for quality assessment in Dutch intensive care.
References


