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.
Summary
Summary

Due to several developments, such as budgetary constraints, insurance regulations and professional ambitions quality assessment of health care is getting increased attention. To enable assessment of the quality of care we need data. Several initiatives have therefore led to the setting up of medical registries. The value of these registries strongly depends on the quality of the data contained in them. The general objective of the study described in this thesis is to evaluate and to assure the quality of the documentation of data in medical registries and the quality of the processing of data for health care performance measurement.

In this thesis we focus on quality assessment 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 the intensive care population.

**Chapter 2** describes the development of a framework of procedures for data quality assurance in medical registries. The framework was based on a literature review and a case study of data quality in the NICE registry. Procedures in the framework were divided into procedures for the coordinating center of the registry (central) and procedures for the centers where the data are collected (local). These central and local procedures were further subdivided into (a) the prevention of insufficient data quality, (b) the detection of imperfect data and their causes (c) actions to be taken / corrections. The framework can be used when a new registry is set up or to identify procedures in existing registries that need to be adjusted in order to improve data quality.

In **chapter 3** we evaluated the effect on data quality of one of the procedures in the framework of chapter 2: training data collectors in data definitions and data extraction guidelines. A training group of 31 Intensive Care physicians extracted data from three sample patient records before and after attending a training program. A control group of ten Intensive Care physicians extracted the data twice, without training in between. In the training group the mean percentage accurate data increased significantly after training (+7% (95% CI 5-10)). The percentage data error due to non-adherence to data definitions decreased by 3.5% after training. Data accuracy in the control group did not change between the two data extractions and was equal to post-training data accuracy in the training group. We concluded that training in data definitions and data extraction guidelines is an effective way to improve quality of intensive care scoring data.
In chapter 4 we assessed data quality in the NICE registry, after implementation of quality improving procedures. Data from a random sample of admissions to nine Dutch ICUs were re-abstracted from the paper patient record or the Patient Data Management System and compared to the data contained in the registry. The re-abstracted data were considered to be the gold standard. The mean percentages inaccurate and incomplete data were 6.1% ± 4.4 (SD) and 2.7% ± 4.4 (SD) respectively. The mean differences in severity-of-illness scores and in predicted mortalities between the registry and the gold standard were very small. We concluded that the current data quality of the NICE registry is good and justifies evaluative research. These positive results might be explained by the implementation of several quality assurance procedures in the NICE registry.

Chapter 5 describes the assessment of the accuracy and reliability of the Sequential Organ Failure Assessment (SOFA) score. Twenty intensive care physicians each scored 15 patient cases. The intraclass correlation coefficient was 0.889 for the total SOFA score. The weighted kappa values were moderate (0.552) for the central nervous system, good (0.634) for the respiratory system and almost perfect (>0.8) for the other organ systems. To assess accuracy the physicians’ scores were compared to a gold standard based on consensus of two experts. The total SOFA score was accurate in 53% (n =158) of the cases. The mean of the absolute deviations of the recorded total SOFA scores from the gold standard total SOFA scores was 0.82. Common causes of errors were inattention, calculation errors and misinterpretation of scoring rules. The results of this study indicate that SOFA scoring among physicians is reliable and accurate. We did advise to implement additional measures to further improve reliability and accuracy.

In chapter 6 we used literature to obtain an overview of aspects related to the content of terminological systems and of methods for the evaluation of the content of terminological systems. Of all methods described in literature three were selected: (1) Concept matching in which two samples of concepts representing respectively (a) documentation of reasons for admission in daily care practice and (b) aggregation of patient groups for research, are looked up in the terminological system in order to assess its coverage; (2) Formal algorithmic evaluation in which reasoning on the formally represented content (Description Logics) is used to detect inconsistencies; and (3) Expert review in which a random sample of concepts are checked for incorrect and incomplete terms and relations. These evaluation methods were applied in a case study on the terminological system DICE. None of the methods applied in the case study covered all the aspects of the content of a terminological system. The results of concept matching differed for the two use cases (63% vs. 52% perfect matches).
Expert review revealed many more errors and incompleteness than formal algorithmic evaluation. We concluded that to get good insight into the content of a terminological system, using a combination of evaluation methods is preferable.

**Chapter 7** describes the validation of four prognostic models (SAPS II, APACHE II, MPM\(_0\) II and MPM\(_{24}\) II) on the NICE registry database, and the analysis of the effects of sample size on measured performance. For each model discrimination (Area Under the Curve (AUC)) and accuracy (mean squared error (MSE) and cross-entropy) statistics, and three calibration statistics (Hosmer-Lemeshow H and Ĉ, and Copas Z) were measured on data from 41,239 ICU admissions. We simulated the validation process with smaller datasets (n = 100, 250, 500, 750, 1000, 2500, 5000), randomly drawn from the large database. The random selection process and the validation were repeated 500 times for each sample size. Differences in performance between models, except MPM\(_0\) II, are small. The AUC, MSE and cross-entropy showed large variation with small sample sizes. The averages of these three statistics were not influenced by the sample size. The applied calibration statistics did appear to be influenced by the size of the validation sample. With larger sample sizes the calibration statistics increased and lack-of-fit appeared more frequently. When using a small validation dataset, it appears to be largely a matter of chance which of the four models outperforms the others. Based on these results validation of prognostic models should not be performed on datasets smaller than 2500 observations.

In **chapter 8** the main results and their implications, and recommendations for future research are discussed. The framework for data quality assurance provides a good reference for existing and coming registries. However, its actual value and usability still have to be proven. Continuous reviewing and maintenance is also a prerequisite for terminological system’s content. The methods for evaluation of terminological systems applied in our case study on DICE should be use in combination with each other. Other methods, which were not applied in our case study, should also be analyzed. Reliable assessment of quality of care is not ensured by high data quality alone but also by the validity of the data processing. Our study provides valuable recommendations for future validation studies of prognostic models.