QUALITY ASSESSMENT OF INPUT DATA FOR EMERGENCY DEPARTMENT SIMULATION

L. Vanbrabant[(a)], N. Martin[(b)], K. Ramaekers[(c)], K. Braekers[(d)]

[(a),(c),(d)] UHasselt, Research group logistics, Agoralaan, 3590 Diepenbeek, Belgium
[(b)] UHasselt, Research group business informatics, Agoralaan, 3590 Diepenbeek, Belgium
[(a),(d)] Research Foundation Flanders (FWO), Egmontstraat 5, 1000 Brussel, Belgium

[(a)]lien.vanbrabant@uhasselt.be, [(b)]niels.martin@uhasselt.be, [(c)]katrien.ramaekers@uhasselt.be, [(d)]kris.braekers@uhasselt.be

ABSTRACT

Emergency Departments (EDs) constitute an important component in a healthcare system. Recently, they are confronted with a substantial growth in demand. Combined with the ever tightening budgets, this has led to the problem of overcrowding in many EDs. Simulation has been widely used in operations management research for analysing and improving patient flow in EDs. The quality of input data is of great importance to build a realistic simulation model. In this paper, data quality problems in healthcare records of emergency departments are identified based on a case study in a Belgian university hospital. The problems are categorised and data quality assessment techniques are developed for each category. A combination of quantitative and qualitative metrics is described to estimate the potential impact of the data quality issues on simulation.

Keywords: data quality problems, data quality assessment, simulation, emergency departments, electronic health records

1. INTRODUCTION

Emergency Departments (EDs) constitute an important component in a healthcare system. They are one of the main entry points of a hospital, offering non-stop healthcare services to patients with various needs. From a social point of view, it is crucial that EDs work efficiently, since timely and good services can save lives. However, EDs are large, complex and dynamic units which are difficult to manage. Moreover, EDs are confronted with a substantial growth in demand due to the ageing population and the trend toward utilising the ED for non-emergency care. Combined with the ever tightening budgets, this has led to the problem of (over)crowding in many EDs. Overcrowding occurs when the demand for emergency services exceeds the available resources in the ED (Bergs et al. 2016, Carmen and Van Nieuwenhuyse 2014).

Currently, ED overcrowding is considered a major international problem. It has significant consequences for both patients and caregivers (Bergs et al. 2016). A lack of sufficient resources prevents timely and suitable services, leading to increased length of stay of patients, increased waiting times, patient dissatisfaction, increased probability of patients leaving the ED without treatment and increased stress levels of caregivers. To face these challenges hospital managers are continuously exploring opportunities to improve the efficiency of their healthcare system without reducing the quality of care (Ahmed and Alkhamis 2009, Carmen and Van Nieuwenhuyse 2014).

Operations Research and Operations Management (OR/OM) techniques have been widely applied to analyse and optimise processes in healthcare organisations (e.g. Saghafian et al 2015). Since EDs are complex and stochastic systems, leading to stochastic outputs, the complete system cannot be modelled analytically and the stochastic outputs can only be evaluated through simulation. Simulation also makes it possible to investigate the simultaneous effect of different improvements. In this way, the simulation model can take interdependencies into account. Moreover, it is possible to analyse and optimise different measures of emergency department performance.

The first step in a simulation analysis is to build a realistic simulation model. In this respect, two key issues have to be considered that have an impact on the extent to which the model reflects reality. First of all, patient flow through the ED results from the interplay of many factors, so modelling the ED as a whole gives a more realistic view. Most simulation models of an ED focus on the treatment phase, while patient flow through an ED consists of three phases: inflow, throughput (or treatment) and outflow (Asplin et al. 2003, Saghafian et al. 2015). The inflow part is the arrival process in the ED. Arrivals are either by ambulance or by patient walk-in. The treatment part consists of triage, registration, placement in an ED bed, clinical assessment, treatment and diagnostic testing. The last part of patient flow, the outflow, is the disposition process. A patient can be discharged, kept under observation or admitted to an inpatient unit (Carmen and Van Nieuwenhuyse 2014, Gul and Guneri 2015). Modelling all three parts makes a simulation model more realistic, but only if there is sufficient and error-free information available for all three phases. Therefore, the second important issue is the quality of the data used as input to the simulation model. The Garbage-In Garbage-Out principle states that the

ISBN 978-88-97999-89-8; Bruzzone, Frascio, Longo and Novak Eds.

input data used has a direct effect on the quality of process analysis and improvement (Mans et al. 2015, Oliveira et al. 2005).

Data acquisition, data quality assessment and data quality improvement are three necessary steps preceding the construction of a simulation model. Data can be acquired through interviews, observations, surveys or electronic health records (EHRs). Previous research on simulation in EDs does not take data quality into account or lacks a description of the data cleaning process. This paper focuses on data quality assessment of input data extracted from the electronic health records (EHRs) of an emergency department. Since EHRs are frequently used as input data in ED simulation studies, there is a need for a structured approach in assessing the quality of this data. The purpose of this paper is to clarify the problem and importance of data quality in operations research. Based on a case study in a Belgian university hospital, data quality problems faced in the EHRs of an ED are identified and a framework for categorising these problems is developed. A combination of quantitative and qualitative measures is proposed to assess the extent of the data quality problems in each category. The framework in combination with the assessment methods provides guidance to researchers for inspecting input data before use. The data acquisition process and quality problems with regard to database development and improvement are beyond the scope of this paper.

2. PROBLEM CONTEXT

This paper is based on data extracted from the EHRs of the ED of a Belgian university hospital. The hospital under study is confronted with ED overcrowding, caused by an increase in the number of patient visits without a proportionate capacity expansion. The total number of visits to the ED approximated 57.650 in 2016 and is expected to increase in 2017. As simulation is an effective tool for the analysis and improvement of ED operations (Oh et al. 2016, Saghafian et al. 2015), the final goal is to build a realistic simulation model of the ED. The extracted data file will be used as input to this model. The file contains anonymised patient records for all patients that visited the ED in November 2016, December 2016 and January 2017. The first step, before building the simulation model, is to assess the quality of this data file as input to the simulation model.

EHRs are used throughout the entire hospital to standardise data gathering and to facilitate data exchange between departments. The software used for the EHRs captures the medical information of every patient and his flow throughout the hospital. Each patient has its own record with a unique patient number in the database. Patient records in the ED contain personal information, mostly obtained by read-in of the identity card. Furthermore, medical and patient flow information is registered at every stage in the ED. This information contains, amongst others, symptoms, diagnosis, type of inflow, timestamps of the patient flow through the ED, outflow destination, etc. Some data is gathered automatically due to triggers in the system, e.g., if a CT-scan is ordered, a timestamp of the order is automatically added to the patient record. Other information has to be inserted manually by a physician, nurse or administrative clerk such as the triage code of patients and the diagnosis. EHR data registration is a process in which individuals with a wide range of backgrounds, all working in the ED, are involved. They all attach different importance to data registration and the precision of the data inserted into the system (Kahn et al. 2012). Additionally, data registration is not the primary focus of healthcare providers. This makes the intrinsic quality of data in EHRs questionable. In assessing the quality of the extracted data file, the primary focus of this paper is on its suitability as input for the simulation model. The data has to be qualitative enough for reuse in the operations research domain. Otherwise, the results of the research can be misleading and of little value. The fitness for use concept indicates that data can be suitable for one research area or for one type of stakeholders, but of low quality for another. Patient data are recorded for operational and managerial purposes inside the hospital (e.g. monthly overviews using scorecards and personnel assessments) and for clinical research. They are not gathered with a focus on reuse in the operations management domain (Kahn et al. 2012, Wang and Strong 1996, Weiskopf and Weng 2012).

Data quality assessment is essential to appraise the intrinsic quality and fitness for use of the extract from the EHRs as input data for the simulation model. Data quality assessment is preceded by the identification of potential data quality problems. If problems are identified, their extent and impact can be assessed by using quantitative metrics and expert judgement (Kahn et al. 2012, Pipino et al. 2002)

Based on the dataset on one hand and on-field observations and interviews on the other hand, data quality problems present in the EHRs of the ED are identified in this paper. The focus lies on data quality problems with a potential impact on simulation results. To this end, quality issues concerning the input data required in the simulation model are the main focus. The input data needed for the simulation model depend on the software and process model used. The simulation model will be built in Arena, a discrete-event simulation software provided by Rockwell Automation. Some of the necessary input data in the Arena software include: durations of service times, patient arrival times, patient categories and processing rules, resources and their capacity etc. (Guo 2016). Concerning the process model, patient flow can be divided in three stages: inflow, throughput and outflow. For the simulation model to be a good reflection of reality, all three stages have to be included at a desirable level of detail (Asplin et al. 2003, Saghafian et al. 2015).

3. DATA QUALITY PROBLEMS

An overview of the attributes included in the dataset under study is provided in Table 1. Within this dataset, several data quality problems can be distinguished. Firstly, certain attribute values are not recorded for all

patients. Timestamps of the first consultation by a physician, the first time a patient is assigned to a box in the ED and the time a patient is medically finished (i.e. approved by a physician to leave the ED) are some examples. Another attribute that is missing for some patiens, is the triage code and a timestamp of the triage process. The triage code indicates the severity of a patient's symptoms. The triage process is only executed between 7 a.m. and 10 p.m., so most missing values are due to the fact that patients arriving during the night shift were not subjected to the triage process. The missing attribute value is not really a quality problem, but the triage code is a necessary input value to the simulation model because patient streams and service times are determined according to triage code.

Secondly, some patient records contain implausible attribute values. First, timestamps may not follow the logical order of patient flow throughout the ED. Consider a patient for which the timestamp of the first consultation with a physician falls before the timestamp of triage or a patient which is only medically finished after leaving the ED. Second, mutually dependent activities take place separately sometimes. An example of this quality issue is the fact that a radiological or laboratory examination request is ordered, but the examination never started and no results are received. Also, a mutation request (i.e. admission request to an inpatient unit) and plan do not always precede an admission to an inpatient unit. Third, attribute values can be incorrect or imprecise without being incoherent with other attribute values. Medical staff sometimes bundles administrative tasks for a group of patients, so timestamps do not always reflect the exact time of an activity. Typing mistakes are another common source of incorrect values.

Finally, particular attributes can be absent in the data file. If a fundamental input variable for the simulation model is missing, the quality of the results is doubtful. Sometimes the values of these attributes can be derived from other, known, attributes, but these tend to be approximations which build upon particular assumptions. Some missing attributes in the data file of the hospital under study are the end times of activities, which are needed to calculate service durations. Other examples include the resources carrying out an activity and the different types of radiological examinations that a patient has undergone.

The aforementioned problems are some examples of quality problems that might be present in the EHRs of an ED. There are several reasons underlying these problems. The ones most commonly indicated by medical and administrative staff are described below. First of all, medical staff has other priorities and can forget to register actions at busy moments. They also indicated that rules exist with reference to patient flow. However, those rules are not always complied with. Context, situation, experience and gut feeling play a role in the decision making process in an ED. For example, a child that is very upset and has the same triage code as an adult, but a later arrival time, can be treated earlier. Another potential reason of data quality problems, is the fact that some units within the ED work independently. These units are radiology, psychiatry, paediatrics and the laboratory. They have their own resources, EHR system and practices. Integrating data of all units within the ED can create inconsistencies. Furthermore, records of patients leaving the ED to the operating room or intensive care unit may contain quality problems because the primary focus is on saving the patient's life. A last source of data quality problems that is commonly indicated by hospital staff are registration errors (e.g. typing mistakes), since a lot of information is recorded manually.

Table 1: Overview of Attributes in the Data File under Study

ATTRIBUTE	EXPLANATION
TIMESTAMPS (Date + time)	
Start date (only date)	The date a patient arrives at the ED and is first registered in the system
T Arrival	Timestamp expressing patient registration in the system
T First triage	Timestamp representing the completion of triage (the point at which a triage code is entered in the system)
T First physician	Timestamp at which the doctor starts writing a report after a first consultation with the patient
T First physical location other than waiting room	Timestamp when the patient was moved out of the waiting room to another physical location (box) for the first time
T Start observation	Timestamp at which the doctor decides that the patient needs to be placed in observation
T Medically finished	Timestamp at which the doctor "signs off" the patient (all medical actions are completed from the perspective of the ED)
T Mutation request	Timestamp when a bed in the hospital was requested for the patient
T Mutation plan	Timestamp when a bed in the hospital is assigned to the patient
T Departure	Timestamp when patient left the ED
T Rx request	Timestamp of the first request for a radiological examination (entered by the physician)
T Rx start execution	Timestamp when the radiological examinations are executed
T Rx first report	Timestamp of the first finished report of the radiological examinations
T Rx last report	Timestamp of the last finished report of the radiological examinations
T Lab request	Timestamp of the first request for a lab test (blood, urine, …)
T Lab first sample received	Timestamp when the first sample is taken for a lab test
T Lab first report	Timestamp when the first finished report was written of the lab results
T Lab last report	Timestamp when the last finished report was written of the lab results
T Pharmacy first use	Timestamp when something was taken from the electronic medicine cabinet (eg. medication, band aid, ...)
T Last triage	Timestamp when the final triage code was given
NUMERICAL	
Patient number	Unique number assigned to every patient, used for identification purposes
File number	Unique number for every file available for a patient e.g. every time a patient visits the hospital, a new file is opened
Age	The age of the patient
Discharged outside the hospital	Dummy variable indicating if a patient is discharged to a place outside the hospital e.g. home, other hospital, nursing home…
CATEGORICAL	
First triage code	The first triage code assigned to a patient (ESI-triage, code between 1-5)
Last triage code	The last triage code assigned to a patient
Mutation unit	The inpatient unit an admitted patient is assigned to
Brought by	Indicates if a patient came to the hospital by ambulance, police, walk-in, transfer or internal transport
Destination after ED	Indicates the destination of the patient after the ED, for example home, inpatient unit, nursing home, other hospital, passed away…
Discharge type	A patient can be discharged on medical advice, admitted, LWBS, left against medical advice or passed away.
FREE TEXT	
Main complaint	Most important symptoms of a patient when arriving in the ED
Diagnosis	The final diagnosis made by a physician, registered at the time of departure. This should be a categorical attribute, but the ICD-9 coding is not consistently followed.

4. DATA QUALITY FRAMEWORK

The previous section outlined potential data quality problems. In order to identify such problems, thorough quality investigation of the data recorded in EDs is required. This matter receives limited attention in literature on ED simulation. Consequently, there is a need for a structured approach for evaluating the quality of data from EHRs. In this section, a categorisation of the

problems occurring in the EHRs of an ED is developed, based on existing data quality literature.

4.1. Literature review

In the literature, several general taxonomies for data quality problems have been provided. Table 2 gives an overview of these frameworks and the main classification basis used to categorise data quality problems. Data quality problems can be classified according to granularity level, schema or instance level, problem manifestation and fitness for use.

Table 2: Overview of Existing Data Quality Frameworks and the Main Classification used.

	MAIN CLASSIFICATION			
FRAMEWORK	Granularity level	Schema or instance level	Problem manifestation	Fit for use model
Wang and Strong (1996)				x
Rahm and Do (2000)	x	x		
Kim et al. (2003)			x	
Mueller and Freytag (2003)			x	
Barateiro and Galhardas (2005)	x	x		
Oliveira et al. (2005)	x			
Gschwandtner et al. (2012)	x			
Kahn et al. (2012)				x
Weiskopf and Weng (2012)				x
Mans et al. (2015)			x	

One of the first frameworks was proposed by Wang and Strong (1996). This framework is based on quality aspects that are important to data consumers. It is built around the concept of fitness for use, which emphasises the importance of taking the viewpoint of the end user into account. The framework consists of four dimensions: intrinsic, contextual, representational and accessibility data quality. The first dimension comprises quality problems that are inherent to the data. The second dimension captures the fit for use concept. Data can be accurate, but not of good quality for the application. The last two dimensions are related to the system used for data gathering.

Rahm and Do (2000) created a data quality framework based on two distinctions: (i) single-source vs. multi-source problems and (ii) schema level vs. instance level problems. Single-source problems are concerned with only one dataset and multi-source problems with the integration of multiple datasets. Schema level problems contain data quality issues emerging because of a poor data model design and a lack of enforcement of data entry rules. Instance level problems are data quality problems inherent to the data values. This category is comparable with the intrinsic data quality category of Wang and Strong (1996). The categories of data quality defined by Barateiro and Galhardas (2005) are based on the same distinctions as Rahm and Do (2000). Oliveira et al. (2005) distinguishes four granularity levels based on the different relations apparent in a relational database. This division is comparable with the single- and multiple-source classification, the only difference is that Oliveira et al. (2005) focus on the number of datasets to integrate. Gschwandtner et al. (2012) classify time-oriented data quality problems into single- and multiple source problems.

In other frameworks, the main categorisation is based on the possible data anomalies instead of the granularity level of the data. Mueller and Freytag (2003) divide data quality problems into syntactical anomalies, semantic anomalies and coverage anomalies. All categories are applicable at different levels in a database, from a single dataset to a complete relational database. Kim et al. (2003) developed a comprehensive classification of dirty data based on the manifestation of the quality problem. The main subdivision is between missing and not-missing data. Not-missing data is broken down further into wrong data and not wrong, but unusable data. In all categories, problems present in a single- and multi-source dataset and at the system and instance level can be found.

The main classifier differs between the existing frameworks, but most frameworks overlap in the final data problems identified. In some frameworks these final problems are very specific, so that they can be measured by specific tests (e.g. Barateiro and Galhardas (2005), Gschwandtner et al. (2012), Kim et al. (2003), Oliveira et al. (2005), Rahm and Do (2000)). Examples of those final problem types are missing values, spelling errors, duplicated records, values outside domain ranges etc. Other frameworks define non-overlapping, but broad problem categories, like accuracy, completeness, believability, timeliness, etc. (e.g. Wang and Strong (1996)).

The previous frameworks are general data quality frameworks, applicable and adjustable to nearly every research context. Focusing on data quality in healthcare, three frameworks have recently been developed. These are indicated in grey in Table 2. Mans et al. (2015) define four classes of problem types: missing data, incorrect data, imprecise data and irrelevant data. These problem classes are identified based on event logs from EHRs. An event log is an ordered list of events. An event represents "something" that happens within a process and is related to a case such as a patient. Consider, for instance, the start of an examination for a particular patient. Additional information that can be recorded about the event includes its timestamp and the resource that is associated to the event (Mans et al. 2015). Within the ED context, event logs can convey insights in, for instance, the order in which a patient undergoes activities, the resource executing these activities and, potentially, even on the patient's condition. This information can be highly relevant for simulation purposes. Kahn et al. (2012) and Weiskopf and Weng (2012) classify EHR data quality problems based on the framework of Wang and Strong (1996). The framework is adjusted to only incorporate data quality problems relevant in a healthcare context and especially in the reuse of data for clinical research. As a result, only intrinsic and contextual data quality problems are taken into account.

Since the approach used in the development of data quality frameworks for EHRs focuses on the reuse of data in clinical research, there are still deficiencies with

regard to the use in operations research contexts in general and simulation in particular. Also, the level of detail in the data quality dimensions is insufficient. The categories are too general, with every category still containing a lot of distinct problems. To be able to define quality assessment methods and to avoid overlooking problems that are not immediately recognisable, a greater level of detail is necessary.

4.2. Data quality framework

All discussed frameworks contain particular data quality problems that are present in the dataset of the ED under study, but none of them completely covers all the identified data quality problems. Since insights from the existing frameworks – both general and healthcare specific – are valuable, these form the basis for establishing a new data quality framework.

The objective of this paper is to build a framework to identify data quality problems in an extracted data file of an ED intended for use in an operations research context, especially simulation. We assume the hospital is the only authorised user of the database, so they compose a data file with the requested information. This file is made available to the researcher. This means that a classification based on granularity level is superfluous. Also, since we are not concerned with the design of the data gathering system, only instance level problems are identified. The two remaining classifications, a distinction based on problem manifestation and fit for use models, are both applicable. Since the context is already defined, we decide to use a problem manifestation classification. The framework is developed with a focus on one application domain, but this does not preclude the use in other research contexts such as other operations research studies in healthcare.

The established framework can be found in Figure 1. The main classification used in the framework is based on the framework of Kim et al. (2003), because this fits the extracted data file better than the classification of Mueller and Freytag (2003). The problem classes of Mans et al. (2015) are also covered in the framework. Data quality problems are split into missing data and not-missing data. The latter category is further divided into wrong data and not wrong but not directly usable data. The name of the last subcategory is changed compared to the framework of Kim et al. (2003), where this category is named not wrong but unusable. With regard to the intended use, the new name covers the category's content better. It contains data that is not wrong, but further data processing is required to make it usable for the purpose at hand. A general example is the presence of start- and end timestamps of an activity, while activity durations are needed.

The main classification is further divided until specific data quality problems are identified. This makes it possible to define measures to assess the extent of the data quality problem for every end category of the framework. End categories consists of only one problem type and there is no overlap between them. However, it is possible that a specific problem in a dataset can be classified in multiple categories. Especially if the data quality assessment techniques of more than one category lend themselves to detect the problem. The different categories of the framework are described in sections 4.2.1 to 4.2.3. The numbers between brackets in Figure 1 are used to refer to the structure of the framework.

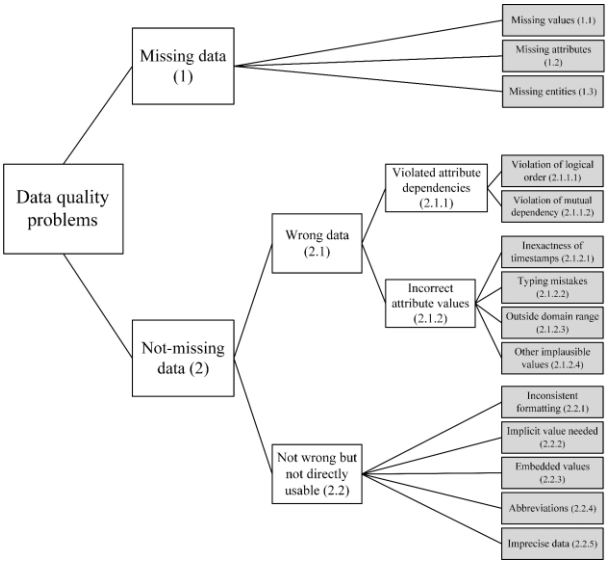

Figure 1: Data Quality Framework for EHRs of EDs in operations research context

4.2.1. Missing data

Missing data (1) is data that is missing in a field while it should not be missing (Kim et al. 2003). Missing data are a very common and inevitable problem (Penny and Atkinson 2001). The fact that some data values are missing can have two important negative effects. First, it can lead to biased estimates for statistics such as central tendency, dispersion or correlation. In a simulation context, biased input parameters can result from missing data. The extent of the negative effect depends on the cause of missing data, i.e. whether missingness is caused by other factors. In case it is related to the (unknown) value of the attribute itself or another attribute in the dataset, it can result in a distortion of the estimates. In case the missing values are randomly distributed in the dataset, the bias is minimal. Secondly, missing data reduce the statistical power of the analysis, because there are less cases available for the analysis (Tsikriktsis 2005). Because missing data can have an impact on the credibility of the simulation study, this is the first category of data quality problems in our framework.

There are three types of missing data: values, attributes and entities. Missing values (1.1) are mandatory attribute values that are missing for certain patients. For example, the triage code is missing for a patient, while triage is executed for every patient arriving at daytime. Other examples are timestamps of performed activities and the discharge type of a patient.

Missing attributes (1.2) are attributes needed as input to the simulation study that are not recorded in the data file. The difference with the previous category is that the values of these attributes are missing for every patient in

ISBN 978-88-97999-89-8; Bruzzone, Frascio, Longo and Novak Eds.

the dataset. Sometimes the attributes are recorded in the EHRs but not included in the extracted data file. Another possibility is that the attributes are not recorded at all. It is possible that timestamps of certain activities are missing or that it is unknown which radiological examinations patients have undergone or which resource executed a task.
The last type of missing data are missing entities (1.3). Normally, every arriving patient and every action executed on that patient has to be registered. However, the dataset at hand showed periods in which no patients arrived for extended periods of time. This is not realistic, so there are patients missing in the data file. A possible reason is a technical failure of the system or an error in the data extraction process.

4.2.2. Wrong data

Wrong data (2.1) is the first of two not-missing data categories. Quality problems manifesting themselves as wrong data are grouped into violated attribute dependencies and incorrect attribute values.
Violated attribute dependencies (2.1.1) are data values that cannot be identified as wrong without information about other attribute values. The violation of logical order category (2.1.1.1) describes problems with the timestamps of successive activities. For example, a patient can only be triaged after arrival, radiological examinations are executed after a first consultation by a physician and no actions can happen to the patient after he left the ED.
The second type of problem related to attribute dependencies is a violation of mutual dependency (2.1.1.2). Attributes are mutually dependent if the value of one attribute affects the value of another attribute. An example is the fact that a patient who has been admitted to the hospital, needs to have a mutation request and mutation plan timestamp and an internal unit assigned to him. Other examples are that if a patient has never seen a physician, his discharge type has to be set at 'left without being seen' and that a patient aged under 16 will be seen by a paediatrician.
Incorrect attribute values (2.1.2) are data values that are wrong on their own, without violating their relation with other attributes. This category contains four problem types: inexactness of timestamps, typing mistakes, values outside domain ranges and other implausible values. The first problem type (2.1.2.1) indicates the fact that timestamps may be recorded imprecisely. Physicians giving low priority to administrative tasks, sometimes results in bundling these tasks for several patients. The timestamps are an inaccurate representation of the activity time because the registration is done afterwards. Also, timestamps can be wrong because of input mistakes if they are not acquired automatically at the time a doctor changes a medical file.
The second problem type are typing mistakes (2.1.2.2), e.g. a typing mistake in the diagnosis field. The focus in this category is on text fields, because typing mistakes in numerical or categorical fields may be identified in one of the other subcategories of the incorrect attribute values class, or they may be unidentifiable (e.g. triage code 3 instead of 4 is registered). Also, typing mistakes are very clear in text fields because it leads to inexistent words, but in numerical fields they are more difficult to identify. In numerical or timestamp fields, typing mistakes manifest themselves as values outside the domain range or implausible/inexact values. Assigning these errors to typing mistakes is difficult, so we do not consider them in this category.
Values outside the domain range (2.1.2.3) are the third problem type. This category includes timestamps, numerical and categorical values that are impossible given the domain ranges. A timestamp has to lie between the start and end of the data extraction period, triage codes have to be values between 1 and 5 and there are five possible discharge types for a patient, namely discharged home, admitted to the hospital, left against medical advice, left without being seen and passed away.
The last problem type is a residual category for wrong data values that do not fit in one of the previous ones (2.1.2.4). For example, resource information can be wrong if a resource forgets to log out from the system, so every action on a computer is registered as done by the same resource. This makes it seem like an implausible number of actions are executed by one resource within a certain time period.

4.2.3. Not wrong but not directly usable data

The second not-missing data quality category (2.2) is different from the previous one in that the data values are not wrong. However, the raw data is not suitable for the specific task at hand. After some data processing efforts, the values can still be used in the analysis. This category contains five specific problem types: inconsistent formatting, implicit value needed, embedded values, abbreviations and imprecise data. Inconsistent formatting (2.2.1) means that there is an inconsistency in the coding of the values within one attribute or among attributes. There are several possibilities: the same representation can be used for different values (e.g. an empty field indicates either a zero or a missing value), different representations for the same value (e.g. a zero is indicated by 0 or an empty field) and a different format for the same value types (e.g. the diagnosis is coded with ICD-9 or free text, dates are presented as DD-MM-YY or YY-MM-DD).
Implicit value needed (2.2.2) means that there is no value present for a patient because an action is not executed or not all details of the action are registered in the data file. Since the attribute is inherent to each patient or activity, this value can be assigned without executing the process or registrating all activity details. If the value is needed in the simulation study, this is perceived as a quality problem. The fact that triage is not executed at night, while patient flow through the ED depends on triage code, fits in this category. Even though patients arriving at night have a particular severity of their condition and, hence, an implicit triage code, no explicit value will be assigned. The presence of start- and end times of an activity, while the duration is needed, can also be

categorised as implicit value needed. The value is implicitly present in the data file, but not recorded as a separate attribute.
Embedded values (2.2.3) are the third problem type, indicating data fields containing more than one value. For example, a timestamp field may contain date and time information, while only time information is needed to create an arrival distribution depending on the hour of the day.
Abbreviations (2.2.4), the fourth problem type, are also correct values, but their meaning has to be derived to be useful. Finally, the imprecise data category (2.2.5) comprises values that are correct but do not contain the necessary amount of detail. For example, it is indicated that radiological examinations are executed but not which specific examinations. Another problem fitting in this category are timestamps with only a date of execution, not the exact time.

5. DATA QUALITY ASSESSMENT

The data quality framework gives an overview of possible quality issues in an ED dataset. Data quality assessment techniques can be used to check the presence of a certain problem type. Also, the severity of the quality problem can be quantified for a number of problem types. To this end, possible techniques for identifying and measuring data quality problems are provided for the end categories present in the framework. These categories are indicated in grey in figure 1.

5.1. Missing data

5.1.1. Missing values

The presence and quantity of missing values seems straightforward to identify, but an important consideration has to be made. In case null values are not possible for an attribute, every empty, n.a. or zero field indicates a missing value. So for mandatory attributes, the number of missing values is easily determined by counting the number of missing values. If missing values are not consistently represented, all representations have to be defined before counting. Since absolute values have no meaning without a reference value, the percentage of missing values for a specific attribute i can be calculated as follows:

$$\frac{Total\ number\ of\ missing\ values\ for\ attribute_i}{Total\ number\ of\ records\ in\ data\ file} * 100\% \quad (1)$$

In the other case, if null values are possible for an attribute, a distinction has to be made between missing and null values. A possible way to do this is by identifying dependencies with other attributes. For example, if a patient's discharge type is 'left without being seen', the timestamp of the first consultation with a physician is not recorded. In all other cases, this timestamp has to be present. Another example is that a mutation plan and mutation request are not assigned for a patient who is discharged home, otherwise this value is missing. By identifying certain dependencies between the attribute under study and other attributes in the data file, missing values can be identified. After that, the number of missing values can be counted and the extent of the problem can be determined by formula (1).
An important consequence of missing values is the existence of incomplete records. A lot of incomplete records undermine the possibility to reconstruct the exact patient flow through the ED. So besides assessing the missingness for every attribute separately, the amount of complete patient records is also an important measure. Since missing values do not necessarily occur within the same patient records for different attributes, the amount of incomplete patient records is not just the maximum of formula (1) over all attributes. Instead, in the most extreme case, it can be the sum of formula (1) over all attributes. To calculate the amount of incomplete records, every patient record has to be checked for missing values in one of the attributes. The percentage of incomplete records is defined with formula (2):

$$\frac{Total\ number\ of\ incomplete\ records}{Total\ number\ of\ records\ in\ data\ file} * 100\% \quad (2)$$

5.1.2. Missing attributes

Missing values are relatively easy to identify and quantify in comparison with missing attributes and entities. Regarding missing attributes, the number of missing attributes depends on the application. In case of simulation, the necessary attributes depend on the specific part of the ED to model and on the amount of detail taken into account. If attributes are missing, the severity of this quality problem is contingent on the derivability of the attribute values from other data or the possibility to deduce a good estimate based on on-field observations or surveys. Given these considerations, it is possible to assess the presence of this quality problem, but measuring the extent of the problem is a subjective evaluation by the user of the data given the specific application.

5.1.3. Missing entities

Concerning missing entities, it is also difficult to quantify the problem. Since missing entities are not registered, the number of missing entities is not deductible from the data file. However, it is possible to determine if the quality problem is present, since it is characterised by extended time periods without arrivals. Those time periods are longer than the normal interarrival times. The maximum possible interarrival time is based on judgment by ED personnel.

5.2. Wrong data

5.2.1. Violation of logical order

Violation of logical order implies that the patient flow based on timestamps in the data file is not correct compared to the normal patient flow. Since patient flow is site- and context-specific, the first step is to define the order of the n events in the regular patient flow:

$$T\ event_1 < T\ event_2 < \cdots < T\ event_n \quad (3)$$

Note that some activities can take place in parallel or random order, so these activities have to be excluded from the order of events (e.g. different examinations).
Only timestamps of sequential activities have to be checked, since this automatically implies that all other dependencies are satisfied:

$$T\ event_i < T\ event_{i+1} \quad \forall i < n \tag{4}$$

The extent of the quality problem can be measured by calculating the number of records for which the logical flow of events is violated and dividing it by the total number of records.

$$\frac{Number\ of\ records\ with\ T\ event_i > T\ event_{i+1}}{Total\ number\ of\ records\ (excl.\ missing\ T)} * 100\% \quad \forall i < n \tag{5}$$

The records with missing values for one or more timestamps are excluded from the denominator, since it is impossible to define the patient flow of those patients based on incomplete records. If patient flow depends on patient characteristics, caution should be exercised, since there is more than one possible ordering of events. When violation is suspected based on the regular patient flow, but the recorded patient flow seems possible given patient characteristics, expert confirmation is advisable.

5.2.2. Violation of mutual dependency

Mutual dependency means that the value of one attribute has an impact on the value of another attribute. For example, a patient with an age below 16 has to be assigned to paediatrics. If this dependency is violated, it indicates an error in one of the attribute values. To assess the data quality for this problem type, the number of records for which a specific mutual dependency is broken, has to be calculated. This number is divided by the total number of records in the data file. Records with missing values for the attributes under study are not excluded, since the presence of a value for one attribute can imply that the other attribute also has to be present or vice versa. Therefore, the fact that a value is present or missing can also imply a violation of mutual dependency.

$$\frac{Number\ of\ records\ with\ attribute_i \nsim attribute_j}{Total\ number\ of\ records\ in\ data\ file} * 100\% \tag{6}$$

An important note with this formula is that only attribute couples (i,j) with a mutual dependency between attributes i and j are tested on this quality issue.

5.2.3. Inexactness of timestamps

This category contains timestamps that are possible and hence do not violate any dependency with other attributes, but they are not realistic. By calculating KPI's or durations, outliers can be identified. These outliers can indicate incorrect timestamps for the attributes used in the calculations. Examples are length of stay (T departure – T arrival), door to doctor time (T first physician – T arrival) or durations between successive events. The percentage of inexact timestamps can be calculated by formula (7):

$$\frac{Total\ number\ of\ outliers}{Total\ number\ of\ records\ (excl.missing)} * 100\% \tag{7}$$

The denominator does not contain records with missing values for one of the attributes used as input to the calculations. An outlier can be identified as a record for which the absolute standardised value is larger than 4 (see Hair et al. 2009 for more information). Given the unpredictable and complex nature of an ED, it is difficult to identify which derived values can be seen as outliers. Even though outlier analysis can be used to identify possible quality problems, the measures have to be interpreted carefully. Resource information can also be an indication of inexact timestamps. If the number of activities executed by one resource at more or less the same time is unrealistic, there is a high probability that the resource bundled the administrative tasks for several patients.

5.2.4. Typing mistake

The focus of this quality issue is on text fields with typing mistakes. This means that a typing mistake can be identified as an unknown word by using a list of possible words given the attribute (e.g. diagnosis) or a dictionary. Since this is very complex, quality assessment for this problem type will not be discussed further.

5.2.5. Outside domain range

Numerical attributes (or timestamps) with a value smaller than the minimum or larger than the maximum acceptable value lie outside the domain range. Correct attribute values meet the following equation:

$$Min_{Ai} \leq Value_{\ Ai} \leq Max_{Ai} \tag{8}$$

The subscript Ai stands for attribute i. For categorical attributes, this quality problems exists if the assigned value is no element of the possible value set. Normally,

$$Value_{Ai} \in \{value\ set\} \tag{9}$$

To assess the quality for this problem type, formula (10) can be used.

$$\frac{Number\ of\ values\ outside\ domain\ range\ attribute_i}{Total\ number\ of\ records\ (excl.missing)} * 100\% \tag{10}$$

In this equation, the denominator contains all entities with a value recorded for the attribute under study. Missing values are excluded since we cannot check the domain ranges.

5.2.6. Other implausible values

Since this is the rest category, it is not possible to define a general assessment method for these problems. If there is a suspicion of a quality problem not captured in the previous categories, this can be checked and a percentage of implausible values can be calculated. An example is calculating the number of actions executed by one resource within a given time period. If this amount is not realistic, it is an indication of incorrect resource

information. Another indication are resources executing tasks outside their shift times. Note that this is not the same as the bundling of administrative tasks, which is already covered in the inexactness of timestamps category.

5.3. Not wrong but not directly usable

For not wrong but not directly usable data, approaches to identify the presence of the problem types are defined. Moreover, solution methods to convert the data to values usable in the application are proposed. The ease of transformation is important to estimate the impact of the data quality problem. As this category does not contain unsolvable problems, quantifying the problem is not of great value

5.3.1. Inconsistent formatting

By inspecting data values of the same type within and among attributes, differences in coding can be identified. If the problem of inconsistent formatting is present, the values of the attributes have to be reformed for a consistent representation. For standard data formats (e.g. dates, times, names…), this can be easily done by changing the cell properties with a data analysis or spreadsheet program (e.g. Microsoft Excel, R). In case of application domain specific coding, like ICD-9 codes in a healthcare context, reformatting is more complex. A possible solution is to look for the ICD-9 code most closely related to the description of the diagnosis. The free text value can be replaced with the associated code.

5.3.2. Implicit value needed

As indicated in the discussion of the data quality framework, missing values have to be separated from null values. Null values occur when an activity has not been executed for a patient. Sometimes, the attribute value is inherent to the patient, so a value can be assigned without executing the activity. If the attribute is a necessary input to the simulation model, the implicit values have to be determined. Consider null values for triage code during the night shift as an example. There are several ways to define the triage code afterwards. Based on the symptoms and diagnosis in the data file, an expert can be asked to define the triage codes for patients arriving at night. Also, observations at night can give an indication of the distribution of patients according to triage code. Another possibility is to estimate the distribution at night based on the daytime distribution if they are similar. The distribution of diagnoses (ICD-9 codes) can also be used as an approximation of the triage code distribution.

This category also contains activity information that is not recorded. The attribute is not missing since the value is included implicitly in other attribute values. An example are service times, which can be derived from start- and end timestamps of an activity. By calculating the difference between the start and end of the activity, the service times can be deducted.

5.3.3. Embedded values

Attributes with problematic embedded values can be identified by first indicating the values needed as input to the simulation model. If these values are present in combination with another value within one field, the attribute value has to be split in different values to be useful. Embedded timestamps are attributes containing date and time, while only the time is valuable. In Excel, time information can be obtained by subtracting the date from the attribute value. In other data editor software, like R, it is possible to split the attribute value according to a prespecified rule.

5.3.4. Abbreviations

Abbreviations can be present in a free text field. An abbreviation is easy to recognise, but detecting attributes containing abbreviations is time-consuming. The most straightforward way to do this is by just going through the data file and checking the attribute values. Standard abbreviations can be simply transformed to complete words, but for domain-specific terminology, expert assistance is desirable.

5.3.5. Imprecise data

The last problem type, imprecise data, is relatively easy to identify, but hard to solve. Measuring units of an attribute value are an indication of precision and they are clear from the data file. If the attribute values do not contain enough detail for the target use, extra data can be extracted if recorded. For example, radiological examinations are only indicated by a date, but timestamps are available in the database of the radiological unit. This information can be requested. Otherwise, empirical data gathering or modelling on a higher abstraction level are possible solutions.

6. CONCLUSIONS AND FUTURE RESEARCH

In this paper, data quality problems and assessment techniques are developed from the viewpoint of simulation in EDs. The reliability of a simulation model developed to analyse ED performance, depends on the input data used. Therefore, the quality of the data used as input to the simulation model has to be investigated. Data acquisition, data quality assessment and data quality improvement are three necessary steps preceding the construction of a simulation model. The focus of this paper is on quality assessment of the EHRs in an ED. Data quality assessment is preceded by the identification of potential data quality problems. The focus is on problems with a possible impact on simulation, so data quality problems related to the input data of the model. Based on an extracted data file from the ED of a Belgian university hospital, a data quality problem framework is proposed. The framework builds upon insights from previous research, with the main classification of Kim et al. (2003) as starting point. Quality problems are divided in missing data and not-missing data problems. Not-missing data has two subcategories: wrong data and not wrong but unusable data. This main classification is further divided until specific data quality problems are

identified. A total of fourteen specific, non-overlapping problem types are found. For each problem type, identification methods are presented together with formulas for quantifying the extent of the quality problem if relevant.

The purpose of this paper was to clarify the problem and importance of data quality in operations research. Not only survey data, but also electronically recorded data can contain errors and should be checked systematically and used carefully. The framework in combination with the assessment methods provide guidance to researchers for inspecting input data before use. As a next step, a tool for data quality assessment will be developed building upon the conceptual foundations outlined in this paper. This tool will require certain context-specific inputs by the researcher, such as the logical flow of events to test violation of logical order. This information will be used to test data files on several quality aspects and to indicate problem areas in the dataset.

There are several possibilities for future research. A possible topic is the extension of the framework to other operations research techniques and to operations research in other healthcare domains. Also, the framework can be tested and eventually modified for use in other countries to enhance the generalisability. Another interesting field to investigate more thoroughly, is the development of quality assessment techniques for the different data quality problems identified in the framework. Some basic techniques are provided in this paper, but there may be more advanced possibilities. Finally, investigating improvement techniques for the different data quality problems is a valuable direction for future work.

REFERENCES

Ahmed M.A., Alkhamis T.M., 2009. Simulation-optimization for an emergency department healthcare unit in Kuwait. European Journal of Operational Research, 198, 936-942.

Asplin B.R., Magid D.J., Rhodes K.V., Solberg L.I., Lurie N., Camargo C.A., 2003. A conceptual model of emergency department crowding. Annals of Emergency Medicine, 42(2), 173-180.

Barateiro J., Galhardas H., 2005. A survey of data quality tools. Datenbank-Spektrum, 14, 15-21.

Bergs J., Vandijck D., Hoogmartens O., Heerinckx P., Van Sassenbroeck D., Depaire B., Marneffe W., Verelst S., 2016. Emergency department crowding: Time to shift the paradigm from predicting and controlling to analysing and managing. International Emergency Nursing, 24, 74-77.

Carmen R., Van Nieuwenhuyse I., 2014. Improving patient flow in emergency departments with OR techniques : a literature overview. FEB Research Report KBI_1425, KU Leuven, Belgium.

Gschwandtner T., Gärtner J., Aigner W., Miksch S., 2012. A taxonomy of dirty time-oriented data. International Conference on Availability, Reliability, and Security, 58-72.

Gul M., Guneri A.F., 2015. A comprehensive review of emergency department simulation applications for normal and disaster conditions. Computers & Industrial Engineering, 83, 327-344.

Guo H., Goldsman D., Tsui K., Zhou Y., Wong S., 2016. Using simulation and optimisation to characterise durations of emergency department service times with incomplete data. International Journal of Production Research, 54(21), 6494-6511.

Hair J.F.J., Black W.C., Babin B.J., Anderson R.E., 2009. Cleaning and transforming data. In: Hair J.F.J., Black W.C., Babin B.J., Anderson R.E., 7th ed. Multivariate Data Analysis: a global perspective. Pearson, 33-90.

Kahn M.G., Raebel M.A., Glanz J.M., Riedlinger K., Steiner J.F., 2012. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Medical care, 50(7), 21-29.

Kim W., Choi B.J., Hong E.K., Kim S.K., Lee D., 2003. A taxonomy of dirty data. Data Mining and Knowledge Discovery, 7, 81-99.

Mans R.S., van der Aalst W.M.P., Vanwersch R.J.B., 2015. Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. Switzerland:Springer.

Oh C., Novotny A.M., Carter P.L., Ready R.K., Campbell D.D., Leckie M.C., 2016. Use of a simulation-based decision support tool to improve emergency department throughput. Operations Research for Health Care, 9, 29-39.

Oliveira P., Rodrigues F., Henriques P.R., 2005. A Formal Definition of Data Quality Problems. Proceedings of the International Conference on Information Quality.

Penny K.I., Atkinson I., 2012. Approaches for dealing with missing data in health care studies. Journal of Clinical Nursing, 21, 2722-2729.

Pipino L.L., Lee Y.W., Wang R.Y., 2002. Data quality assessment. Communications of the ACM, 45(4), 211-218.

Rahm E., Do H.H., 2000. Data cleaning: problems and current approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.

Saghafian S., Austin G., Traub S.J., 2015. Operations research/management contributions to emergency department patient flow optimization : review and research prospects. IIE Transactions on Healthcare Systems Engineering 5(2), 101-123.

Tsikriktsis N., 2005. A review of techniques for treating missing data in OM survey research. Journal of Operations Management, 24, 53-62.

Wang R.Y., Strong D.M., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, 12(4), 5-33.

Weiskopf N.G., Weng C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151.