Building a Foundation for High-Quality Health Data: A Multi-Hospital Study in Belgium

Jens Declerck, Bert Vandenberk, Mieke Deschepper, Kirsten Colpaert, Lieselot Cool, Jens Goemaere, Mona Bové, Frank Staelens, Koen De Meester, Eva Verbeke, Elke Smits, Cami De Decker, Nicky Van Der Vekens, Elin Pauwels, Robert Vander Stichele, Dipak Kalra, Pascal Coorevits

Submitted to: JMIR Medical Informatics
on: May 06, 2024

Table of Contents

Building a Foundation for High-Quality Health Data: A Multi-Hospital Study in Belgium

Jens Declerck[1, 2] MSc; Bert Vandenberk[3] MD, PhD; Mieke Deschepper[4] PhD; Kirsten Colpaert[4] Prof Dr; Lieselot Cool[5] PhD; Jens Goemaere[5]; Mona Bové[6] PhD; Frank Staelens[6]; Koen De Meester[7] MSc; Eva Verbeke[7] IR; Elke Smits[8]; Cami De Decker[8]; Nicky Van Der Vekens[9] DVM, PhD; Elin Pauwels[10] PhD; Robert Vander Stichele[11] Prof Dr; Dipak Kalra[1, 2] Prof Dr; Pascal Coorevits[1] Prof Dr

[1]Department of Public Health and Primary Care Unit of Medical Informatics and Statistics Ghent University Ghent BE
[2]The European Institute for Innovation through Health Data Ghent BE
[3]Department of Cardiovascular Sciences University Hospitals Leuven Leuven BE
[4]Data Science Institute Ghent University Hospital Ghent BE
[5]Data Science AZ Groeninge Kortrijk BE
[6]Department of Quality and Process Managenemt OLV Hospital Aalst BE
[7]Cell Business Intelligence AZ Sint-Lucas Ghent BE
[8]Clinical Research Center Antwerp University Hospital Antwerp BE
[9]Clinical Data Manager General Hospital Maria Middelares Ghent BE
[10]Quality Department General Hospital Maria Middelares Ghent BE
[11]Faculty of Medicine and Health Sciences Heymans Institute of Pharmacology Ghent BE

Corresponding Author:
Jens Declerck MSc
Department of Public Health and Primary Care
Unit of Medical Informatics and Statistics
Ghent University
Campus UZ-Ghent, Entrance 42, 6th Floor
Corneel Heymanslaan 10
Ghent
BE

Abstract

Background: Data quality is fundamental to maintain the trust and reliability of health data for both primary and secondary purposes. However, before secondary use of health data, it is essential to assess the quality at the source and to develop systematic methods for the assessment of important data quality dimensions.

Objective: This case study aims to offer a dual aim: to assess the data quality of height and weight measurements across seven Belgian hospitals and to outline the obstacles these hospitals face in sharing and improving data quality standards

Methods: Focusing on data quality dimensions completeness and consistency, this study examined height and weight data collected from 2021 to 2022 within three distinct departments – surgical, geriatrics, and paediatrics – in each of the seven hospitals.

Results: Variability was observed in the completeness scores for height across hospitals and departments, especially within surgical and geriatric wards. In contrast, weight data uniformly achieved high completeness scores. Notably, the consistency of height and weight data recording was uniformly high across all departments.

Conclusions: A collective collaboration among Belgian hospitals, transcending network affiliations, was formed to conduct this data quality assessment. This study demonstrates the potential for improving data quality across healthcare organizations by sharing knowledge and good practices, establishing a foundation for future, similar research.

(JMIR Preprints 06/05/2024:60244)
DOI: https://doi.org/10.2196/preprints.60244

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Building a Foundation for High-Quality Health Data: A Multi-Hospital Study in Belgium

Abstract

Background: Data quality is fundamental to maintain the trust and reliability of health data for both primary and secondary purposes. However, before secondary use of health data, it is essential to assess the quality at the source and to develop systematic methods for the assessment of important data quality dimensions.

Objective: This case study aims to offer a dual aim: to assess the data quality of height and weight measurements across seven Belgian hospitals and to outline the obstacles these hospitals face in sharing and improving data quality standards.

Methods: Focusing on data quality dimensions completeness and consistency, this study examined height and weight data collected from 2021 to 2022 within three distinct departments – surgical, geriatrics, and pediatrics – in each of the seven hospitals.

Results: Variability was observed in the completeness scores for height across hospitals and departments, especially within surgical and geriatric wards. In contrast, weight data uniformly achieved high completeness scores. Notably, the consistency of height and weight data recording was uniformly high across all departments.

Conclusions: A collective collaboration among Belgian hospitals, transcending network affiliations, was formed to conduct this data quality assessment. This study demonstrates the potential for improving data quality across healthcare organizations by sharing knowledge and good practices, establishing a foundation for future, similar research.

Introduction

In an era of digital healthcare, hospitals are collecting structured and coded electronic health record (EHR) data at high velocity resulting in a high volume of potential valuable data [1, 2]. The secondary use of this data specifically provides the opportunity to accelerate research and improve patient care pathways [3]. Belgium, already a major hub for clinical trials in Europe [4], is enhancing its reputation through the Real-World Data for Belgium (RWD4BE) [5] initiative, which collaborates with stakeholders to improve healthcare by reusing patient data for both primary and secondary purposes.

However, amongst these initiatives and opportunities that the evolving healthcare landscape presents, the quality of health data is imperative. Data quality is the cornerstone ensuring the trustworthiness and reliability of health data use and reuse. Despite its importance for safe patient level care and accurate inferences, obtaining high quality data in a healthcare setting remains a challenge, accompanied with ambiguities in defining data quality and the most suitable assessment methods [6-8].

A widely accepted definition in the literature is that of data being “fit for purpose” [9]. This concept is further refined by The Global Data Management Community (DAMA international), that adapts this definition to a more specific data context: “data quality is the degree to which the data dimensions meet requirements”[10].This definition expands upon earlier ones by capturing the subjectivity and context-dependence inherent in data quality, offering a more stakeholder-sensitive perspective that aligns with the specificities of data use and reuse.

The path for ensuring data quality in secondary use of health data is complex and multifaceted. It involves not only the original quality of the data when and where it is captured (e.g., within the primary source such as an EHR system) but also the quality of the processes by which the data is transferred and transformed for further use, for example by mapping the data to a data model and terminology systems used within a clinical data warehouse. These stages, each embedded within the comprehensive data lifecycle, are often overlooked in the literature [8], yet they are essential for a thorough understanding of data quality.

Prior to making secondary use of health data, it is crucial to measure the quality at the source and to establish methodologies for assessing relevant data quality dimensions. If not, it will not be possible to recheck the quality of the data at later life-cycle points to verify that it has not been degraded (e.g., through an Extract, Transformation, Load (ETL) process). Research into data quality, especially when involving multiple primary data sources, consistently encounters significant challenges that add to the complexity of the research process [11]. These include measurement discrepancies [12], the use of varied software systems for data collection [13], inconsistent coding of diseases and procedures [14, 15], and complex data sharing agreements [16]. Together, these factors not only hinder the efficient exchange of data but also significantly affect the quality of the data [17]. Furthermore, the healthcare ecosystem faces a significant challenge due to the lack of clear and practical guidelines for implementing strategies to ensure a high data quality, especially when sharing data across different health care organizations for secondary use.

Several studies have aimed to define data quality dimensions and methodologies to describe and measure the dynamic complexity of data quality [6, 18-21]. Despite these efforts, there is still no comprehensive framework that captures all aspects of data quality [8]. This has led to a fragmented understanding of data quality dimensions, with varying interpretations depending on the specific use of health data. Literature suggests that existing methods are often constrained by the absence of standardized metrics that can accurately assess data quality across different dimensions [8]. This limitation also extends to the transformation of these dimensions into concrete requirements for primary and secondary data usage, as well as for the extract, transformation, load (ETL) processes, that consider the original intent being the data's collection at the primary source.

This paper examines these challenges by presenting a case study on data quality across multiple Belgian hospitals. The study is driven by a twofold objective: firstly, to evaluate the data quality within these hospitals, focusing specifically on the critical example variables of height and weight; and secondly, to identify and document the challenges and obstacles hospitals face when implementing and managing data quality improvement initiatives.

Methods

In this case study the data quality framework developed by the European Institute for Innovation through Health data (i~HD) was adopted [8, 22]. This data quality framework is the prior result of analysing 22 different published frameworks and consolidating all of their defined data quality dimensions into a consolidated framework, condensed into nine distinct data quality dimensions. Table 1 presents the data quality dimensions incorporated into the i~HD data quality framework together with their definitions. As completeness and consistency dimensions were the most frequently used in the data quality literature, these two were selected for quality assessment in this study[8, 23].

Data quality dimension	Definition
Completeness	The extent to which data is present
Consistency	The extent to which data satisfies constraints
Correctness	The extent to which data are true and unbiased
Timeliness	The extent to which data is promptly processed and up-to-date
Stability	The extent to which data are comparable among sources and over time
Contextualisation	The extent to which data are annotated with acquisition context
Representativeness	The extent to which data are representative of intended use
Trustworthiness	The extent to which data can be trusted based on the owner's reputation
Uniqueness	The extent to which data are not duplicated

Table 1 i~HD Data Quality Framework

Case study

Study setting

The case study was conducted within seven different hospitals across Belgium. Supplementary Table 1 presents a detailed table of the bed capacity for each participating hospital reflecting its scale and patient intake capability. To protect the confidentiality of the participating hospitals, they are referred to using numerical identifiers such as hospital 1, hospital 2, etc. This approach safeguards the privacy of all participating hospitals.

The variables of interest in this case study were height and weight. These two basic variables are both crucial health metrics that inform a wide range of clinical decisions [24-26]. These variables were used to exemplify the quality of data that should be reliably captured in all patient encounters, thereby serving as a barometer for overall quality of data collection practices within a healthcare setting. To accomplish the objectives set in this article we focused on three specific departments within each hospital, coded with the ward identifier or department code C (surgical), G (geriatrics), and E (pediatrics). The use of these identifiers allows a standardized approach for comparing health data within similar specialty departments across the different hospitals.

The mentioned departments were not selected at random but based on their distinct patient care characteristics and the importance of data quality related to their specialized care. Height and weight are important for clinical care decision making in all three specialties. Surgery often involves immediate and precise measurements for dosing anesthetics and ventilation settings during the procedures [27], geriatrics requires attention to detail given the complex health profiles and frailty of elderly patients [28], and pediatrics involves growth monitoring, which is vital for assessing child and adolescent health [29, 30].

Data collection

The data collection spanned a two-year period and included all patients discharged from the respective departments between January 1, 2021, and December 31, 2022. For every patient discharged within this timeframe, the last recorded data for height and weight were extracted. Should there be instances where measurements for these variables were not taken, these were systematically recorded as ‘NA’ or ‘NULL’ to denote the absence of data within the specified timeframe. It is important to note that measurements taken during pre-admission consultations were deliberately excluded from the data collection. This decision was made to ensure that the study focused on inpatient data, reflecting the data quality at this point of care during hospital stays. Since natural language processing falls outside the scope of this case study, only structured data was extracted. Supplementary table 2 presents the total number of patients by department for each hospital involved in this case study.

The research team’s access to and analysis of the data was limited to aggregated results, for data protection reasons. Only hospital personnel with the appropriate access rights performed the data extraction and derived the data quality assessment results. The aggregated results were derived from the data quality assessments conducted within each hospital using a standardized set of analytical tools in R (version 2023/09/1) [31]. This approach not only facilitated the uniform assessment of data quality across various institutions but also eliminated the need for direct access by the research team to personal health information, thereby preserving patient anonymity and confidentiality within this research.

Data for this study was collected in accordance with data sharing agreements established with all participating hospitals (University of Leuven, University Hospital Gent, General Hospital Groeninge, OLV Hospital Aalst, General Hospital – AZ Sint-Lucas Gent, Antwerp University Hospital, General Hospital Maria Middelares) and the study protocol was approved by the Ethics Committee of Antwerp University Hospital (Project ID 6268). The challenges and obstacles encountered by hospitals were highlighted during an interactive conference session. During this session, various hospitals shared their experiences, and participants offered their insights and feedback [32-35].

Data quality assessment and statistical analysis

The first step in the data quality assessment focused on assessing the completeness of the collected data. Completeness, in this case study, refers to the extent to which height and weight data were recorded for patient admissions within the study's time frame. Completeness was quantified as the percentage of missing values for both height and weight data across all selected departments.

The second step evaluated the consistency of the data. In this case study, consistency was determined by the percentage of recorded height and weight values falling within clinically acceptable ranges, predefined based on department-specific norms. To calculate consistency scores, missing data entries were excluded to avoid skewing the results. Subsequently, the proportion of data within the specified ranges for both height and weight were then determined, providing a percentage score of values falling inside and outside the acceptable range. Table 2 provides an overview of the variables and associated data quality rules for completeness and consistency.

Variable	Definition	Departments	Data quality rule
Height	Height (m)	Geriatrics (code G)	Range between 1,4 – 2,2 meters
		Surgical (code C)	
		Paediatrics (code E)	Range between 0,4 – 2 meters
Weight	Weight (kg)	Geriatrics (code G)	Range between 40 – 160 kg
		Surgical (code C)	
		Paediatrics (code E)	Range between 1 – 80 kg

Table 2 Overview of variables and data quality rules used for consistency

Results

Data quality assessment

Data quality assessment for completeness in hospitals

Figure 1 presents all completeness (%) scores for height and weight for each hospital across the described departments. Each boxplot contains the interquartile range (IQR) of the completeness percentage, with the central line within each box representing the median value. The square box presents the average value. Supplementary table 2 compiles all results for each department within every hospital.

In the geriatrics department, the mean completeness of height data stood at 63,89%, with a median marginally lower at 62,35%. A wide range was observed, stretching from 25,56% to 94,70%, reflecting significant variability among institutions within this department, as evidenced by an IQR of 31,77%. Conversely, the pediatrics department presented a mean completeness of 60,11% with an even larger IQR of 73,10%, suggesting a more pronounced discrepancy in recording practices. Here, data completeness varied from a minimum 11,29% to a maximum of 97,26%. The median value, at 77,24%, was higher than the mean, indicating a distribution skewed towards lower completeness percentages. The surgical department reported a mean completeness of 63,12% for height data, with an IQR of 51,75% underscoring the variability. The lowest recorded completeness was 33,92%, and the highest was 96,99%, with a notably lower median of 44,67% suggesting a skew towards less complete data.

Figure 1. Completeness scores for height and weight data

A comparison with weight data showed different trends. The geriatrics department exhibited a high mean completeness of 92,73% and a relatively narrow IQR of 7,62% indicative of more uniform data collection practices. Completeness ranged from a high 77,97% to an exemplary 98,72%, with the median at 96,74%, pointing to a cluster of values towards the upper range. The pediatrics department's mean completeness for weight data was 89,94%, with an IQR of 6,06%, denoting consistency. Despite some outliers, as the range spanned from 52,24% to 99,09%, the elevated median of 97,53% implied that most pediatric units adhered to high standards of data completeness. In the surgical department, the mean completeness was 87,63%, with a modest IQR of 3,56%, reflecting uniformity in data capture. With a range from 46,43% to 97,94% and a median of 93,79%, the findings suggested that while data recording is generally robust, there is room for improvement.

Data quality assessment for consistency in hospitals

Figure 2 presents all consistency (%) scores for height and weight for each hospital across the described departments. Each boxplot contains the interquartile range (IQR) of the consistency percentage, with the central line within each box representing the median value. The square box presents the average value. Supplementary table 2 compiles all results for each department within every hospital.

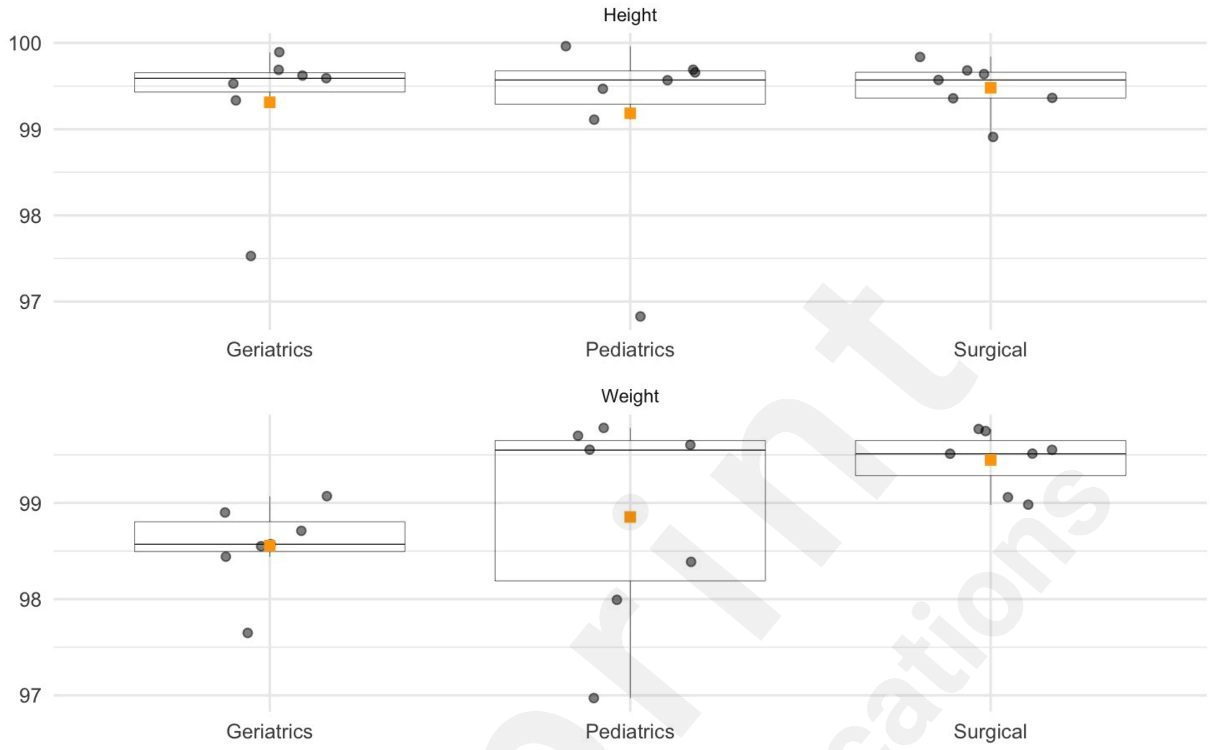

Figure 2. Consistency scores for height and weight data.

Height data recording demonstrated high consistency, with the geriatrics department achieving a mean of 99,31%. The consistency was impressively uniform, ranging narrowly from 97,53% to 99,89%, as the median at 99,59% and a minimal IQR of 0,23% confirmed. The pediatrics department’s mean consistency was 99,18%, with a slightly broader range from 96,83% to 99,96%. Nevertheless, a high median of 99,57% and an IQR of 0,39% indicated a strong overall consistency. Similarly, the surgical department showed a mean consistency of 99,48%, with a small range from 98,91% to 99,84%. The median of 99,57% and an IQR of 0,30% denoted a highly reliable level of data quality.

Consistency in weight data also exhibited positive results. In the geriatrics department, the mean consistency was 98,56%, with a tight range between 97,65% and 99,07%. The median mirror the mean, and an IQR of 0,31% highlighted the concentration of values. In pediatrics, the mean consistency was 98,85%, with a wider range from 96,97% to 99,78% and an IQR of 1,46%. Despite this variability, a high median of 99,55% was maintained. The surgical department continued the trend of high mean consistency at 99,45%, with a tight range and an IQR of 0,37%, signaling a consistent quality of weight data recording.

Challenges and obstacles towards data quality initiatives

Privacy/governance requirements

Data quality initiatives within hospitals are complicated by privacy and governance requirements that dictate the handling and sharing of patient data. These regulations are critical for protecting patient confidentiality but pose substantial challenges for collaborative data quality initiatives. An important challenge in this domain is the heterogeneity of documents concerning data sharing agreements. Each hospital, with its unique set of protocols and agreements, operates with its own distinct processes for data sharing. This fragmented approach towards data sharing and governance creates significant difficulties when attempting to share data across different hospitals. Not only does it impede the flow of information, but it also poses a barrier to improving data quality, as harmonizing data across different systems becomes a complex task.

Software requirements

The diversity in hospital software systems, comprising both commercially available and internally developed solutions, poses a considerable challenge to initiatives aimed at ensuring data quality. Although we introduced a standardized protocol, the heterogeneity in these systems utilized by each participating hospital had an impact on the data extraction process. Additionally, while some hospitals extracted data directly from the original source, others retrieved it from their data warehouse. This variance in methods introduces multiple steps in the data extraction phase, each presenting areas for potential data quality issues even before the data quality assessments starts.

Additionally, the variation in data management systems leads to differences in the data quality control protocols when entering data or there might not be designated structured fields for certain variables in the EHR system. Some participating hospitals indicated that although the completeness score was low, data for height and weight were available in the EHR. However, it was often embedded within unstructured fields, limiting its utility for trend analysis, clinical decision support and further complicating data harmonization efforts. All these differences in software tools utilized by the hospitals may lead to subtle yet impactful biases or variations in the data when data is transferred between hospitals.

Who is responsible?

An important challenge encountered in this study was the varying levels of responsibility for data quality across the hospitals. Collaboration with data scientists within each hospital was essential, as their expertise in handling health data was invaluable in acquiring the necessary information for our case study. Only one of the participating hospitals had a data quality manager responsible for data quality initiatives, and providing insights into existing issues, curation practices, and strategies for data quality improvement. However, it is important to note that there is no solid evidence showing that a hospital with a dedicated data quality manager is more likely to achieve higher data quality. The key seems to be a widespread organizational recognition of the importance of data quality, which can be upheld by either an individual or a team. This fragmented role of responsibilities also creates an ambiguity in formal education and training on data quality in the healthcare sector. The participating hospitals indicated they often resort to a ‘do-it-yourself’ approach to tackle data quality, reflecting on an important dedication for obtaining high standards. Yet, this self-reliant method may lead to inconsistencies when data is shared with other entities, highlighting the need for a more uniform approach to data quality management and education.

Discussion

EHRs are repositories of complex data that hold a wide range of multifaceted data that encompass a myriad of patient health metrics, extending well beyond fundamental measurements such as weight and height. This data is intrinsically more complex, bearing significant implications for patient care and health outcomes. Our study and analysis align with other studies in suggesting that the quality of basic variables (e.g. height and weight) may serve as a preliminary barometer for the overarching data quality within EHRs [36-38]. A focused assessment of these metrics can uncover broader trends and highlight the challenges faced in advancing complex data quality initiatives.

Data quality assessment results

The results reveal a distinct difference between completeness and consistency across the departments. For completeness, the results for height and weight varied among hospitals, particularly in the surgical and geriatric departments. The completeness score for height data varied significantly across all different hospitals and departments. This is in contrast with the completeness scores for weight, these scores revealed that most hospitals maintained high standards of data quality with respect to weight records. This difference between data capture indicates the prioritization of the weight variable based on the patient care requirement at each department. The temporal stability of recorded measurements for height and weight also presents notable difference. This inherent stability of height may not prompt repeated measurements during subsequent visits, contributing to the observed variability in the completeness scores for height data. Weight is a more dynamic metric, with potential for significant change between visits. Therefore, weight is documented more systematically due to their critical role in monitoring medication dosage, leading to its higher completeness scores and illustrating the rationale behind the different approaches to capturing these metrics.

The consistency scores for both height and weight measurements demonstrate remarkable uniformity across all departments, suggesting that once recorded, the data is reliably captured. This uniformity in consistency scores underscores a rigorous and systematic approach to data recording when it is performed. However, the ambiguity between these high consistency scores and the more variable completeness scores indicates that while the processes for data entry are robust, the initial act of data capture is not consistently executed. This incompleteness could potentially stem from a high burden of workload or an overload of administrative tasks, which can affect the thoroughness and attention to detail during the data collection phase. This highlights the need for a more standardized approach to the initial data capture, ensuring that all patient records are as complete as they are consistent. Reflecting on these findings, comparative analyses with existing research reveal a consistent pattern of discrepancies in data capture, as well as variations in the results for different variables across multiple studies [39-41]. A key distinction of this study, as opposed to others, lies in the breadth of our assessment across multiple hospitals. Furthermore, a primary goal was to identify pre-existing challenges, which directly influence the data quality assessment and its outcomes.

Challenges and obstacles towards data quality initiatives

This case study has brought to light numerous challenges that must be addressed before conducting a data quality investigation. Resolving these challenges could also enhance the assessment of data quality in the future. It emphasizes the necessity for a more uniform and standardized approach when conducting data quality assessments, which ought to be incorporated as a fundamental component for every RWE data project.

Establishing roles for specialized personnel within hospitals, tasked exclusively with overseeing data quality, would be advantageous. This is particularly true in the context of data sharing among multiple healthcare facilities. Data quality managers could play a pivotal role in bringing key stakeholders together to facilitate data sharing agreements, collaborating with software developers to meet technical requirements during data collection, and safeguard the integrity and consistency of data throughout its lifecycle. They can provide critical perspectives on local data capture processes, ensuring that information is correctly recorded and effectively maintained from the point of origin to its intended use.

To support these data quality managers, a standardized education curriculum focused on data quality principles and practices would be beneficial. Such a curriculum would create a uniform understanding of data quality across different sites, which is instrumental when executing multisite data quality projects. Data quality managers equipped with this education, would be empowered with uniform assessment procedures and strategies. This effort in education and role specialization could significantly improve the quality of data throughout its lifecycle. This is a key goal within QUANTUM [42], a project financed by the EU, which aims to create a label certifying data quality and utility for HealthData@EU.

The insights from this study have highlighted substantial opportunities for advancing data quality practices. Each hospital involved in the study has acknowledged the importance of the results, conforming that the insights gained have offered them practical guidance for enhancing their data management processes. Detailed testimonials from the hospitals, which reinforce this positive sentiment, can be found in supplementary table 3. These statements from the hospitals underscore the practical implications of our research and confirm the shared commitment to continuous improvement in data quality.

Limitations

This study selected two data quality dimensions – completeness and consistency (by range) – while acknowledging the value of the other dimensions within our established framework. The focus on these dimensions reflects their prevalent use within the data quality literature [24-26] and their relevance to clinical data assessment [27-30]. The selection of height and weight alone as variables for analysis cannot fully represent the complex and varied nature of patient records, which include critical elements such as vital signs, laboratory findings, and prescribed medications. Therefore, future studies should include a broader spectrum of quality dimensions and clinical variables. This expansion is necessary to gain a more comprehensive view of data quality, the issues that affect reaching good data quality and strategies to address the multifaceted nature of health data within electronic health records. Furthermore, the study's reliance on aggregated data for the purpose of comparing hospitals introduces a limitation. Each hospital independently conducted their analysis and shared only the aggregated results, which constrained the potential for a comprehensive, cross-institutional evaluation of data quality at the individual patient level. This method limits the ability to identify specific data quality patterns that might only emerge through detailed, patient-level analysis.

Conclusions

This study has underscored the complexities and challenges inherent in assessing and assuring the quality of health data across multiple hospital settings. Our findings indicated a significant variance in the completeness of height and weight data among hospitals, underscoring the need for improved data capture protocols. The high consistency scores within recorded data attest to the precision of documentation when height and weight measurements are recorded. Performing a data quality assessment across different hospitals is a complex process in which multiple challenges need to be addressed before statistical analysis can even start. The challenges identified through this case study, particularly regarding privacy, governance, usage of different software systems, and responsibility for data quality, emphasizes the need for more standardized approaches and specialized roles within the data quality management domain. The establishment of dedicated data quality managers and standardized education could bridge these gaps, enabling more effective and uniform data quality assessments and improvements.

Acknowledgments

We extend our deepest gratitude to the RWD4BE initiative for their instrumental role in fostering collaboration among stakeholders to devise a robust strategy for healthcare data reuse in Belgium. Their commitment to facilitating meetings and bridging connections between participating hospitals has been crucial to our case study. We also commend RWD4BE’s efforts in enhancing the awareness and quality of healthcare data across the ecosystem.

We also wish to express our gratitude to the European Institute for Innovation through Health Data (i~HD) for their guidance and support throughout this journey. We are particularly grateful for the workshop provided during i~HD annual conference in Ghent, Belgium. This workshop contributed to our understanding on the data quality needs and challenges within hospitals.

Supplementary table 1

Participating hospital	**Accredited Beds (N) [43]**
University Hospital Leuven	1908
University Hospital Gent	1049
General Hospital Groeninge	1044
OLV Hospital Aalst	842
General Hospital – AZ Sint-Lucas Gent	779
Antwerp Universtiy Hospital	593
General Hospital Maria Middelares	542

Supplementary table 2

Hospital 1					
	# visits	Weight		Height	
		Completeness	Consistency	Completeness	Consistency
Surgical	22634	96,71%	99,77%	33,92%	99,64%
Paediatrics	6109	98,77%	99,70%	95,58%	99,69%
Geriatrics	6615	98,72%	99,07%	59,47%	99,62%
Hospital 2					
	# visits	Weight		Height	
		Completeness	Consistency	Completeness	Consistency
Surgical	21495	97,94%	99,75%	41,86%	99,84%
Paediatrics	4970	99,09%	99,55%	97,26%	99,96%
Geriatrics	6662	98,57%	98,90%	41,07%	99,89%
Hospital 3					
	# visits	Weight		Height	

		Completeness	Consistency	Completeness	Consistency
Surgical	25880	46,43%	98,98%	39,52%	98,91%
Paediatrics	7797	52,24%	97,99%	31,54%	96,83%
Geriatrics	2531	88,42%	98,57%	25,56%	99,69%
Hospital 4					
	# visits	Weight		Height	
		Completeness	Consistency	Completeness	Consistency
Surgical	24784	91,46%	99,55%	90,82%	99,68%
Paediatrics	3482	89,43%	99,78%	12,84%	99,11%
Geriatrics	4475	77,97%	98,71%	77,07%	99,33%
Hospital 5					
	# visits	Weight		Height	
		Completeness	Consistency	Completeness	Consistency
Surgical	31961	92,59%	99,51%	44,67%	99,36%
Paediatrics	8356	94,89%	98,39%	77,24%	99,66%
Geriatrics	3485	91,79%	98,44%	62,35%	99,59%
Hospital 6					
	# visits	Weight		Height	
		Completeness	Consistency	Completeness	Consistency
Surgical	12905	93,79%	99,51%	96,99%	99,57%
Paediatrics	4090	97,53%	99,60%	11,29%	99,57%
Geriatrics	4391	96,74%	97,65%	87%	99,53%
Hospital 7					
	# visits	Weight		Height	
		Completeness	Consistency	Completeness	Consistency
Surgical	10179	94,46%	99,07%	94,07%	99,36%
Paediatrics	3611	97,66%	96,97%	94,99%	99,47%
Geriatrics	556	96,88%	98,55%	95%	97,53%

Supplementary table 3

Hospital	**Quote**
General Hospital Groeninge	“These results demonstrate the critical urgency to draw attention to the importance of data quality within the hospitals. A thorough data quality analysis should be part of each real-world-data project conducted with or within the hospital.”
OLV Hospital Aalst	“Data quality starts with gaining support from healthcare professionals to enter the right data in the right place. “
Antwerp University Hospital	“A standard uniform data sharing agreement (DSA) is needed. This would ensure that data exchange for similar projects would be less complicated. Regarding the data quality results, it seems useful to perform a second use

	case taking into account parameters other than just weight and height and using additional dimensions for the analysis.

References

1. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Aust Health Rev. 2018 Sep;42(5):568-78. PMID: 29986809. doi: 10.1071/ah17255.
2. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A Survey of the Literature on Unintended Consequences Associated with Health Information Technology: 2014-2015. Yearb Med Inform. 2016 Nov 10(1):13-29. PMID: 27830227. doi: 10.15265/iy-2016-036.
3. Duncan R, Eden R, Woods L, Wong I, Sullivan C. Synthesizing Dimensions of Digital Maturity in Hospitals: Systematic Review. Journal of Medical Internet Research. 2022;24(3):e32994. doi: 10.2196/32994.
4. Deloitte. Belgium as a clinical trial location in Europe. 2022; Available from: https://pharma.be/sites/default/files/2024-01/20240131-belgium-as-a-clinical-trial-location-in-europe-2022-vpublic-sent.pdf.
5. Inovigate. Recommendations on a Real-World Data strategy for Belgium. 2022; Available from: https://www.inovigate.com/media/filer_public/36/dd/36ddcd12-564e-4678-9388-053d7adc6b12/report_recommendations_on_rwd_strategy_for_belgium_final_template_used.pdf.
6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality

assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144-51. PMID: 22733976. doi: 10.1136/amiajnl-2011-000681.
7. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. Western Journal of Nursing Research. 2018;40(5):753-66. PMID: 28322657. doi: 10.1177/0193945916689084.
8. Declerck J KD, Vander Stichele R, Coorevits P. Frameworks, dimensions, definitions of aspects and assessment methods for the appraisal of quality of health data for secondary use: a review of reviews. JMIR Medical Informatics. 2024.
9. Juran JM, Gryna FM, Bingham RS. Quality control handbook: McGraw-Hill New York; 1974.
10. Ehrlinger L, Wöß W. A Survey of Data Quality Measurement and Monitoring Tools. Front Big Data. 2022;5:850611. PMID: 35434611. doi: 10.3389/fdata.2022.850611.
11. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal. 2015. doi: 10.5334/dsj-2015-002.
12. Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. BMC Health Services Research. 2017 2017/11/22;17(1):766. doi: 10.1186/s12913-017-2697-y.
13. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. Perspect Health Inf Manag. 2013;10(Fall):1c. PMID: 24159271.
14. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, et al. Digital Health Data Quality Issues: Systematic Review. J Med Internet Res. 2023 Mar 31;25:e42615. PMID: 37000497. doi: 10.2196/42615.
15. Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: A qualitative study. BMC Health Services Research. 2017 11/22;17. doi: 10.1186/s12913-017-2697-y.
16. Devriendt T, Borry P, Shabani M. Factors that influence data sharing through data sharing platforms: A qualitative study on the views and experiences of cohort holders and platform developers. PLoS One. 2021;16(7):e0254202. PMID: 34214146. doi: 10.1371/journal.pone.0254202.
17. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014 2014/11/05;14(1):1144. doi: 10.1186/1471-2458-14-1144.
18. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244. PMID: 27713905. doi: 10.13063/2327-9214.1244.
19. Liaw ST, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: A literature review. J Am Med Inform Assoc. 2021 Jul 14;28(7):1591-9. PMID: 33496785. doi: 10.1093/jamia/ocaa340.
20. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform. 2013 Jan;82(1):10-24. PMID: 23122633. doi: 10.1016/j.ijmedinf.2012.10.001.
21. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020 Dec 9;27(12):1999-2010. PMID: 33166397. doi: 10.1093/jamia/ocaa245.
22. Aerts H, Kalra D, Saez C, Ramírez-Anguita JM, Mayer M-A, Garcia-Gomez JM, et al. Is the quality of hospital EHR data sufficient to evidence its ICHOM outcomes performance in heart

failure? A pilot evaluation. 2021.
23. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc. 2023 Sep 25;30(10):1730-40. PMID: 37390812. doi: 10.1093/jamia/ocad120.
24. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. Methods Inf Med. 2003;42(1):61-7. PMID: 12695797.
25. Doods J, Botteri F, Dugas M, Fritz F. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials. 2014 2014/01/10;15(1):18. doi: 10.1186/1745-6215-15-18.
26. Selbst SM, Fein JA, Osterhoudt K, Ho W. Medication errors in a pediatric emergency department. Pediatr Emerg Care. 1999 Feb;15(1):1-4. PMID: 10069301. doi: 10.1097/00006565-199902000-00001.
27. Fogagnolo A, Montanaro F, Al-Husinat L, Turrini C, Rauseo M, Mirabella L, et al. Management of Intraoperative Mechanical Ventilation to Prevent Postoperative Complications after General Anesthesia: A Narrative Review. J Clin Med. 2021 Jun 16;10(12). PMID: 34208699. doi: 10.3390/jcm10122656.
28. Travers J, Romero-Ortuno R, Bailey J, Cooney M-T. Delaying and reversing frailty: a systematic review of primary care interventions. British Journal of General Practice. 2019;69(678):e61-e9. doi: 10.3399/bjgp18X700241.
29. Hirata KM, Kang AH, Ramirez GV, Kimata C, Yamamoto LG. Pediatric Weight Errors and Resultant Medication Dosing Errors in the Emergency Department. Pediatric Emergency Care. 2019;35(9):637-42. PMID: 00006565-201909000-00009. doi: 10.1097/pec.0000000000001277.
30. Ward CE, Taylor M, Keeney C, Dorosz E, Wright-Johnson C, Anders J, et al. The Effect of Documenting Patient Weight in Kilograms on Pediatric Medication Dosing Errors in Emergency Medical Services. Prehosp Emerg Care. 2023;27(2):263-8. PMID: 35007470. doi: 10.1080/10903127.2022.2028045.
31. Declerck J. Data Quality Assessment. [GitHub Repository] 2024; Available from: https://github.com/JensDeclerck/Data-Quality-Assessment.git.
32. Bové M. How EHR validation techniques fail to assure trustworthiness within the EHR. i~HD Annual Conference 2023; Ghent, Belgium.
33. Cool L. Creating a momentum for data quality in a general hospital. i~HD Annual Conference 2023; Ghent, Belgium.
34. Deschepper M. The CAMEL in the room: A data quality case on an ICU research database. i~HD Annual Conference 2023; Ghent, Belgium.
35. Van Der Vekens N. The data life cycle: Who, what, where & when? i~HD Annual Conference 2023; Ghent, Belgium.
36. Martin PM. Can we trust electronic health records? The smoking test for commission errors. BMJ Health & Care Informatics. 2018;25(2):105-8. doi: 10.14236/jhi.v25i2.970.
37. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. West J Nurs Res. 2018 May;40(5):753-66. PMID: 28322657. doi: 10.1177/0193945916689084.
38. Scholte M, van Dulmen SA, Neeleman-Van der Steen CWM, van der Wees PJ, Nijhuis-van der Sanden MWG, Braspenning J. Data extraction from electronic health records (EHRs) for quality measurement of the physical therapy process: comparison between EHR data and survey data. BMC Medical Informatics and Decision Making. 2016 2016/11/08;16(1):141.

doi: 10.1186/s12911-016-0382-4.

39. Jackson N, Woods J, Watkinson P, Brent A, Peto TEA, Walker AS, et al. The quality of vital signs measurements and value preferences in electronic medical records varies by hospital, specialty, and patient demographics. Sci Rep. 2023 Mar 8;13(1):3858. PMID: 36890179. doi: 10.1038/s41598-023-30691-z.
40. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. Health Information Management Journal. 2021;50(1-2):88-92. PMID: 31805788. doi: 10.1177/1833358319887743.
41. Taye BK, Gezie LD, Atnafu A, Mengiste SA, Tilahun B. Data completeness and consistency in individual medical records of institutional births: retrospective crossectional study from Northwest Ethiopia, 2022. BMC Health Services Research. 2023 2023/10/31;23(1):1189. doi: 10.1186/s12913-023-10127-0.
42. Available from: https://ecrin.org/projects/quantum.
43. Gezondheid Ze. Huidig aantal erkende bedden per vestingsplaats. 2024; Available from: https://www.zorg-en-gezondheid.be/sites/default/files/external/p_v_i_im_020_e002_erkenningssituatie-_hospitalisatiediensten-_huidig_aantal_erkende_bedden_per_vestigingsplaats.pdf.