How to Cite
Alshammari, N. M. S., Alouthah, Mohammad G. S., Alrshidi, Ahmed S. M., Alshammari, Mateb F. N., Alrasheidi, Bander M. H., Alshammari, Mansour F. N., Alsudais, Abdullah S. A., Alsaadi, Hamoud F. F., & Alshammari, Saad N. K. (2018). The reliability of patient information in electronic systems. *International Journal of Health & Medical Sciences*, *1*(1), 47-53. https://doi.org/10.21744/ijhms.v1n1.2272

The Reliability of Patient Information in Electronic Systems

Nawaf Mithqal Saleh Alshammari
KSA, National Guard Health Affairs
Email: alshammarina2@ngha.med.sa

Mohammad Ghatyan Sulaiman Alouthah
KSA, National Guard Health Affairs
Email: alothahmo@ngha.med.sa

Ahmed Saleh Madws Alrshidi
KSA, National Guard Health Affairs
Email: alrasheidiah@ngha.med.sa

Mateb Falah Nahar Alshammari
KSA, National Guard Health Affairs
Email: alshammaryme@ngha.med.sa

Bander Mohammad Haia Alrasheidi
KSA, National Guard Health Affairs
Email: alrashdiba@ngha.med.sa

Mansour Fahad Nasser Alshammari
KSA, National Guard Health Affairs
Email: alshamaryma@ngha.med.sa

Abdullah Sulaiman Abdullah Alsudais
KSA, National Guard Health Affairs
Email: alsudaisab@nGha.med.sa

Hamoud Faraj Freej Alsaadi
KSA, National Guard Health Affairs
Email: alsaadiho@ngha.med.sa

Saad Nghimish Khasram Alshammari
KSA, National Guard Health Affairs
Email: alshemarisa@ngha.med.sa

Abstract---*Electronic health records (EHRs) are digitized data used for clinical research and data science. Data quality in EHRs is crucial for accurate results, including accuracy, completeness, consistency, credibility, timeliness, accessibility, adequacy, comprehensibility, and interpretability. However, the quality of digital data in EHRs is a major concern, with issues such as incompleteness, duplication, poor organization, fragmentation, and insufficient use of coded data. To improve data accuracy, researchers can use statistical measures and validate data using methods like central tendency, dispersion, and goodness-of-fit tests. Missing data in EHRs can be classified into three forms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).*

ISSN 2632-9433
Submitted: 27 April 2018 |Revised: 09 May 2018 | Accepted: 18 June 2018

Challenges in EHR data include documentation without explicit units of measurement, discrepancies in data collection across organizations, and unstructured text data. To ensure high-quality data, researchers should familiarize themselves with the EHR platform, secondary data sources, and data collection methodologies. Regulations at national and international levels are established to ensure the optimal handling, safeguarding, and usage of personal data, including healthcare information. This study provides an overview of the key elements of high-quality electronic health record (EHR) systems, as well as common practices in their deployment. It concludes by discussing the governance and private administration of EHR systems, including specific specifics and comments.
Keywords---*coding systems, data quality, data science, electronic health records, machine learning, review.*

Introduction

Significant volumes of data are regularly documented as a component of the care process. Although the main purpose is to manage patient care for individuals, there are substantial possibilities to use this data to investigate research inquiries of interest. In the United Kingdom, extensive research has been conducted for nearly 25 years using routine primary care data, which has been anonymized at its origin. This data is obtained from the General Practice Research Database, now known as the Clinical Practice Research Datalink (CPRD), as well as other data sources. Additionally, data from multiple practices is combined and linked to specific electronic health record (EHR) systems, such as QResearch and ResearchOne. Anonymized data refers to data in which all identifying elements that can be traced back to the owner are permanently removed.

Alternatively, there are pseudo-anonymization methods that enable the owner to be reidentified through a procedure managed by those responsible for safeguarding the data's security and privacy. Health Data Research UK has established a comprehensive national register of datasets obtained from electronic health records that are accessible for research purposes (Jørgensen et al., 2014). In the Netherlands, a comparable advancement occurred in the early 1990s. The Netherlands Institute for Health Services Research (NIVEL) created the Netherlands Information Network of General Practice (NIVEL-PCD), formerly known as the NIVEL Primary Care Database (Verheij & van der Zee, 2018; Azur et al., 2011; Schweikardt et al., 2016). Additionally, Belgium has its own Intego Network (Schweikardt et al., 2016; Bartholomeeusen et al., 2005).

France has the Système National des Données de Santé (Bezin et al., 2017) and the data warehouse of Assistance Publique-Hôpitaux de Paris (AP-HP). Sweden has a multitude of comprehensive national registers (Ludvigsson et al., 2016). These databases provide vital insights into the use of healthcare services and advancements in population health. In the United States, the use of routine anonymized data has not been a common practice. This is mainly due to the restrictions imposed by the Health Insurance Portability and Accountability Act (HIPAA) regulations, which prohibit the linking of health data from different sources without consent. Additionally, the limited computerization of small office practices has also contributed to the lack of tradition in using such data. However, the primary emphasis has mostly been on collecting and analyzing data from hospital settings, thanks to the support of the National Institutes of Health's Clinical Translational Science Awards (CTSA) (Lee et al., 2014). The use or utilization of administrative data for research purposes is increasingly limited in Europe, mainly due to the implementation of the European General Data Protection Regulation (GDPR) in 2016. Furthermore, data owners are increasingly seeking authority over the use of their data, which poses challenges in the creation of extensive centralized databases (Hamm et al., 2011; Choy et al., 2003).

Data quality in electronic health records (EHR) is a critical aspect to consider. An electronic health record (EHR) is a digitized compilation of a patient's medical information, encompassing essential administrative and clinical data related to their care. This includes demographic details, vital signs, diagnoses, treatment plans, medications, medical history, allergies, immunizations, radiology reports, and laboratory and test results. Electronic Health Records (EHRs) are dynamic and patient-centric documents that provide immediate and secure access to authorized individuals. The use of EHRs aims to enhance the quality of patient care by facilitating the seamless sharing of relevant medical information across various healthcare providers. Meanwhile, the exponential increase in the quantity of Electronic Health Records (EHRs) has generated a surge in curiosity and potential for diverse research endeavours. To guarantee that patients get the necessary treatment and to obtain accurate and dependable research results, it is essential to have high-quality data (Bray & Parkin, 2009; Larsen et al., 2009).

Data quality refers to the overall attributes and qualities of a dataset that impact its capacity to meet the requirements resulting from the intended use of the data (Fundation, 2006). At now, there is no conclusive consensus about the specific elements that determine the quality of data in existing studies. In a study conducted by Feder,

(2018), various aspects of data quality were identified and frequently reported. These components include data accuracy (ensuring that the data is correct and free of errors), completeness (ensuring that the data is sufficient in terms of breadth, depth, and scope for its intended use), consistency (ensuring that the data is presented in a uniform format), credibility (ensuring that the data is considered to be true and trustworthy), and timeliness (ensuring that the data is recorded promptly and used within a reasonable timeframe) (Feder, 2018; Chan et al., 2010; Kahn et al., 2012; Wand & Wang, 1996; Weiskopf & Weng, 2013). Additional dimensions of data quality may encompass accessibility, referring to the availability and ease of retrieval of data, adequacy in terms of the appropriate quantity of data, comprehensibility, ensuring that data is clear and easily understood, interpretability, ensuring that data is presented in a suitable language and units, and other relevant factors.

A multitude of issues were expressed about the quality of digital data in electronic health records (EHRs), including incompleteness, duplication, poor organization, fragmentation, and insufficient use of coded data in EHR processes. According to the well-known programming principle, if you input low-quality or incorrect data, you will get low-quality or incorrect results. Inadequate data quality may have a detrimental effect on the quality of treatment provided to patients, perhaps resulting in long-term harm or even mortality. Furthermore, erroneous data will have a significant influence on public health decision-making. In the following part, we will go more into the difficulties of data correctness and data completeness (Hazen et al., 2014).

The Accuracy of Data

Data accuracy refers to the degree of precision and truthfulness of the data obtained via the Electronic Health Record (EHR) system. Put simply, it refers to the extent to which the information stored in the Electronic Health Record (EHR) properly represents the actual value in the real world. For example, it assesses if the medication list in the EHR correctly displays the correct amount, dosage, and particular pharmaceuticals that a patient is now using. A preliminary investigation assessed the precision of the information in an Australian primary care environment and verified the presence of mistakes and inaccuracies in electronic health records (EHR) (Tse & You, 2011). The pilot research revealed that there were high levels of accuracy in gathering demographic information, while fairly high levels of accuracy were seen in collecting data on allergies and drugs.

There was a significant proportion of unrecorded information as well. Possible causes of data inaccuracy include errors committed by physicians, such as inappropriate use of the "cut and paste" feature in electronic systems (Ozair et al., 2015), as well as errors, loss, or destruction of data during data transmission (Ozair et al., 2015). Methods to enhance data accuracy during collection involve mitigating potential issues with electronic health records (EHR), such as optimizing preference lists, exercising caution when copying data, adapting templates as necessary, and maintaining comprehensive documentation. Additionally, taking proactive measures like conducting periodic internal audits, providing staff training, and keeping a compliance folder can contribute to improved accuracy.

Various methods may be used to evaluate the precision of data (Feder, 2018). Internal validity refers to the comparison of a specific variable inside a dataset with other factors, such as utilizing medicine to validate the illness condition. Internal validation may also be conducted by finding outliers, which are unrealistic numbers such as very high or low blood pressure readings (Bayley et al., 2013). External validity refers to the practice of using several data sources or datasets to verify the correctness of data. For example, if a patient is enrolled in a stroke registry but is reported as not suffering a stroke in the current dataset, this may be cross-checked using alternative data sources. Connecting numerous datasets is often challenging owing to data privacy policies. The researcher can use basic statistical measures to assess if variable values adhere to logical constraints and patterns in the data. These measures include central tendency (such as mean, median, and mode) and dispersion (such as range and standard deviation) for continuous variables and frequencies and proportions for categorical variables. Additionally, goodness-of-fit tests like the Pearson chi-square test can be employed (Feder, 2018). Validation is a process that researchers use to assess the quality of data and detect any flaws that may be present in the data (Bayley et al., 2013).

Data Completeness

Data completeness refers to the extent and characteristics of missing data fields for certain variables or individuals. Typically, these numbers that are not there are referred to as missing data. Data missingness is a prevalent issue in various types of studies. This can restrict the scope of the outcomes being examined, the number of explanatory factors being considered, and even the size of the population being included (Kourou et al., 2015). Consequently, it diminishes the statistical power of a study and generates biased estimates, ultimately resulting in invalid conclusions. Data may be absent or incomplete for a multitude of reasons. Certain data may be omitted as a result of the study's design. For instance, in some surveys, certain questions are only intended for females to respond to, resulting in a

void for guys on that particular topic. Missing data might occur due to the malfunction of certain equipment at specific times. Data may potentially be absent due to participant non-response. Missing data might occur as a result of errors made during the process of collecting or entering data. Therefore, understanding the methods and reasons behind data gaps is crucial for effectively managing and evaluating missing data (Häyrinen et al., 2008; Protti et al., 2009).

Missing data may be classified into three forms based on the underlying reason: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976) (Figure 1). MCAR, or Missing Completely At Random, refers to data that is missing in a way that is unrelated to any other variables or the variable itself. Instances of Missing Completely At Random (MCAR) might include instances when the inability to capture observations is attributed to random malfunctions of experimental devices. The lack of causes is often attributed to external factors that are unrelated to the observations themselves. For Missing Completely At Random (MCAR) data, it is often acceptable to exclude observations that include missing values. The findings will be unbiased, but, the test may lack statistical power due to a decreased sample size. This assumption is implausible and seldom occurs in practical situations (Tsai et al., 2009; Kononenko, 2001).

When data is missing and it is Missing at Random (MAR), the missingness is not random and may be associated with the observed data, but not with the specific value of the variable in question (Sterne et al., 2009). For instance, there is a higher probability that a male participant would not finish a survey on the degree of depression compared to a female participant (Smith, 2008). The absence of data is attributed to gender rather than the severity of depression itself. If we exclude patients with missing data, the findings will be skewed since the majority of the completed observations are from females. Therefore, it is essential to appropriately consider other observable variables of the participants when imputing missing data that are missing at random (MAR).

However, the assumption of MAR is statistically unverifiable and requires further investigation and research. MNAR, or Missing Not At Random, describes instances where the reason for missing data is dependent on the actual value that the missing data would have had. For instance, the participant declines to disclose the extent of their depression due to their profound state of despair. In this scenario, missingness is attributed only to the value itself, and no additional data can be used to anticipate this value. Data that is missing not at random (MNAR) is a greater challenge since it may result in the absence of data from important subgroups. Consequently, this might lead to samples that do not accurately reflect the population of interest. To acquire an impartial estimation of the parameters in this scenario, it is necessary to create a model for the missing data and include it in a more intricate model for predicting the missing values (Kang, 2013).

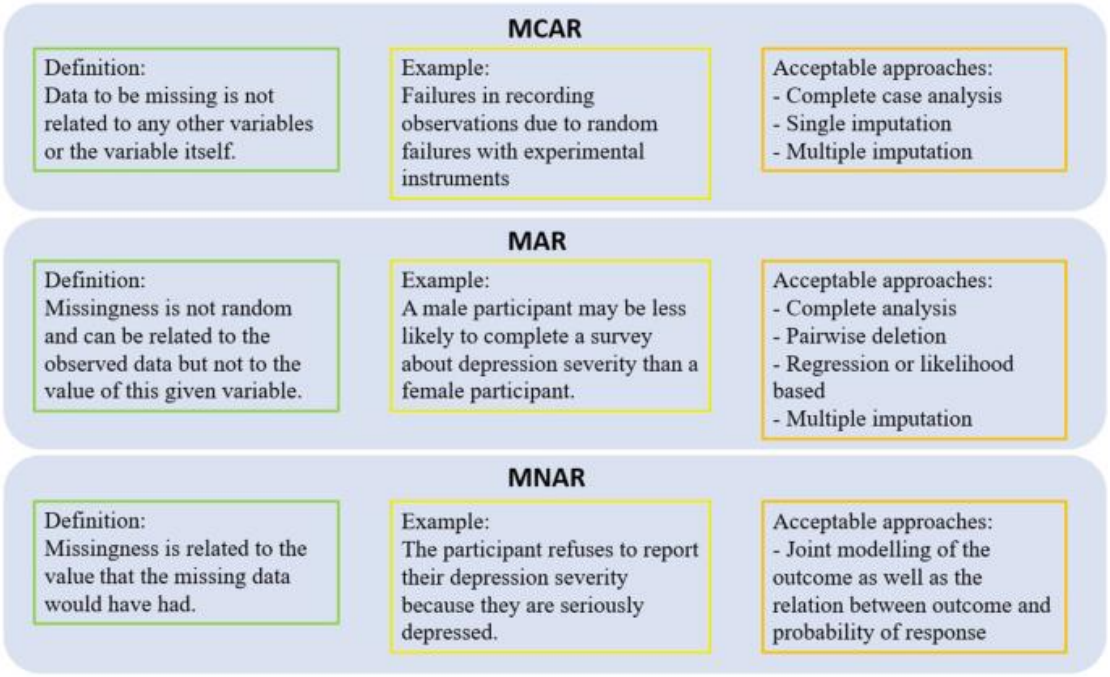

Figure 1. A comprehensive overview of missing mechanisms, with clear definitions, illustrative examples, and recommended strategies for managing missing values

Additional Challenges and Recommendations for Best Practices

Additional obstacles exist in the realm of electronic health record (EHR) data. For instance, some data may be documented without explicitly indicating the units of measurement, hence interpreting such data is challenging (Bayley et al., 2013). Knowing the data-gathering procedure and background information may aid in the interpretation of the data in this scenario. There might be discrepancies in the gathering and categorization of data across different organizations and over the years (Bayley et al., 2013). Certain discrepancies may be readily spotted in the data, such as a metric that was only collected after a certain period. However, some discrepancies could be difficult to detect and need a comprehension of the spatial and temporal aspects of data collection. Lastly, unstructured text data stored in the EHR leads to limited accessibility and other data quality problems, including a lack of impartiality, consistency, and completeness (Bayley et al., 2013). Data extraction methods, such as natural language processing (NLP), are being used to directly identify information from textual notes.

The validity of research outcomes relies on the presence of high-quality data, and the sufficiency of this quality is contingent upon the study's objectives. At now, there are no definitive standards for determining the adequacy of data quality. However, a meticulous examination of data quality should assist researchers in determining if the available data is valuable for the study (Bayley et al., 2013). Feder (2018), advocated three general practices. It is advised to get acquainted with the EHR platform and the secondary data source that is based on EHR. Familiarity with the many categories of data, the methodologies used in data collection, and the identities of the collectors is very advantageous. It is advisable to possess a comprehensive dictionary that provides definitions for all data variables. This dictionary should include information such as the data type, the anticipated range of values for each variable, overall summary statistics, the extent of missing data, and any subcomponents if applicable. The second suggestion is to create a study strategy that incorporates techniques for evaluating and managing the quality of data, such as statistical methods for dealing with missing data and possible measures to address other data quality concerns (e.g., eliminating outliers, and validating diagnostic codes). The final suggestion is to enhance transparency in reporting the quality of data, which includes disclosing the percentage and nature of missing data, any other limitations in data quality, and any modifications made to data values (such as variables excluded from analysis, imputation techniques, variable transformations, and creation of new variables). This will facilitate the use of high-quality data for clinical research. There is a promotion of communication and discussion of the significance of data quality with doctors (Bayley et al., 2013).

Conclusion

The rapidly expanding quantity of electronic health records (EHRs) has generated increased interest and created new possibilities for numerous research endeavors. Data quality is of utmost importance to get relevant and trustworthy study results. The exploration focused on many dimensions of data quality, including data correctness and data completeness, as well as the problems associated with these components. After the data quality section, guidelines for doing data quality analysis were provided.

Regulations at both national and international levels are established to establish fundamental principles that guarantee the optimal handling, safeguarding, and ultimate usage of personal data, including healthcare information. Operationally, there are many hurdles to address, such as the need to anonymize or pseudo-anonymize patients' data and conduct privacy-preserving analysis for both commercial and therapeutic goals. This is especially crucial in machine learning applications since they often need a substantial volume of data. However, sharing data across hospitals is not a feasible or safe approach. Federated learning is currently the most successful and promising technique for implementing machine learning applications in a way that ensures privacy.

References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), 40-49.

Bartholomeeusen, S., Kim, C. Y., Mertens, R., Faes, C., & Buntinx, F. (2005). The denominator in general practice, a new approach from the Intego database. *Family Practice*, *22*(4), 442-447.

Bayley, K. B., Belnap, T., Savitz, L., Masica, A. L., Shah, N., & Fleming, N. S. (2013). Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Medical care*, *51*, S80-S86.

Bezin, J., Duong, M., Lassalle, R., Droz, C., Pariente, A., Blin, P., & Moore, N. (2017). The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiology and drug safety*, *26*(8), 954-962.

Bray, F., & Parkin, D. M. (2009). Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. *European journal of cancer*, *45*(5), 747-755. https://doi.org/10.1016/j.ejca.2008.11.032

Chan, K. S., Fowles, J. B., & Weiner, J. P. (2010). Electronic health records and the reliability and validity of quality measures: a review of the literature. *Medical Care Research and Review*, *67*(5), 503-527.

Choy, K. L., Lee, W. B., & Lo, V. (2003). Design of a case based intelligent supplier relationship management system—the integration of supplier rating system and product coding system. *Expert systems with applications*, *25*(1), 87-100. https://doi.org/10.1016/S0957-4174(03)00009-5

Feder, S. L. (2018). Data quality in electronic health records research: quality domains and assessment methods. *Western journal of nursing research*, *40*(5), 753-766.

Fundation, T.M. (2006), Background issues on data quality. In: The connecting for health common framework

Hamm, J., Kohler, C. G., Gur, R. C., & Verma, R. (2011). Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, *200*(2), 237-256. https://doi.org/10.1016/j.jneumeth.2011.06.023

Häyrinen, K., Saranto, K., & Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, *77*(5), 291-304. https://doi.org/10.1016/j.ijmedinf.2007.09.001

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72-80. https://doi.org/10.1016/j.ijpe.2014.04.018

Jørgensen, A. W., Lundstrøm, L. H., Wetterslev, J., Astrup, A., & Gøtzsche, P. C. (2014). Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial. *PLoS One*, *9*(11), e111964.

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*, S21-S29.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402–406.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

Larsen, I. K., Småstuen, M., Johannesen, T. B., Langmark, F., Parkin, D. M., Bray, F., & Møller, B. (2009). Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. *European journal of cancer*, *45*(7), 1218-1231. https://doi.org/10.1016/j.ejca.2008.10.037

Lee, D., de Keizer, N., Lau, F., & Cornet, R. (2014). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association*, *21*(e1), e11-e19.

Ludvigsson, J. F., Almqvist, C., Bonamy, A. K. E., Ljung, R., Michaëlsson, K., Neovius, M., ... & Ye, W. (2016). Registers of the Swedish total population and their use in medical research. *European journal of epidemiology*, *31*, 125-136.

Ozair, F. F., Jamshed, N., Sharma, A., & Aggarwal, P. (2015). Ethical issues in electronic health records: A general overview. *Perspectives in clinical research*, *6*(2), 73-76.

Protti, D., Johansen, I., & Perez-Torres, F. (2009). Comparing the application of Health Information Technology in primary care in Denmark and Andalucía, Spain. *International journal of medical informatics*, *78*(4), 270-283. https://doi.org/10.1016/j.ijmedinf.2008.08.002

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Schweikardt, C., Verheij, R. A., Donker, G. A., & Coppieters, Y. (2016). The historical development of the Dutch Sentinel General Practice Network from a paper-based into a digital primary care monitoring system. *Journal of Public Health*, *24*, 545-562.

Smith, W. G. (2008). Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior. *Online submission*.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, *338*.

Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. *expert systems with applications*, *36*(10), 11994-12000. https://doi.org/10.1016/j.eswa.2009.05.029

Tse, J., & You, W. (2011). How accurate is the electronic health record?–a pilot study evaluating information accuracy in a primary care setting. In *Health Informatics: The Transformative Power of Innovation* (pp. 158-164). IOS Press.

Verheij, R., & van der Zee, J. (2018). Collecting information in general practice:'just by pressing a single button'?. In *Morbidity, Performance and Quality in Primary Care* (pp. 265-272). CRC Press.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86-95.

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.