MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230949

Understanding Clinician EHR Data Quality for Reuse in Predictive Modelling

Melinda WASSELL[a,1], James L MURRAY[b], Chaithra KUMAR[b], Karin VERSPOOR[c], Kerryn BUTLER-HENDERSON[c]
[a]*Central Queensland University, Australia,* [b]*Work Healthy Australia, Australia,* [c]*RMIT University, Australia*
ORCiD ID: Melinda Wassell https://orcid.org/0000-0002-5171-5271

Abstract. It is imperative to build clinician trust to reuse ever-growing amounts of rich clinical data. Utilising a proprietary, structured electronic health record, we address data quality by assessing the plausibility of chiropractors, physical therapists and osteopaths' data entry to help determine if the data is fit for use in predicting outcomes of work-related musculoskeletal disorders using machine learning. For most variables assessed, individual clinician data entry positively correlated to the clinician group's data entry, indicating data is fit for reuse. However, from the clinician's perspective, there were inconsistencies, which could lead to data mistrust. When assessing data quality in EHR studies, it is crucial to engage clinicians with their deep understanding of EHR use, as improvement suggestions could be made. Clinicians should be considered local knowledge experts.

Keywords. Data quality, electronic health records, machine learning.

1. Introduction

Extensive EHR data collections are essential to inform decisions with real-world evidence. There is an opportunity for machine learning (ML) to help support clinical decision-making and improve the quality of care [1]. The reuse of clinical data to develop ML is challenged by issues with trust [2], specifically due to concerns about data quality [3]. Engaging clinicians who produce data may improve quality and trust.

As EHR research expands, frameworks specific to reusing EHR data are advancing [4, 5]. One parameter within these frameworks that could help clinicians build trust in reuse of data is atemporal plausibility [5], or the believability of the data. Atemporal plausibility examines whether observed data agrees with local knowledge, such as by establishing consistency between clinicians' data input.

For clinicians, spending time on quality data input can take valuable time away from patient interactions. Clinicians using the EHR in this study reported that due to time constraints, they entered varying detail, depending on the severity of the patient presentation, which led to a mistrust in the data quality. The underlying trustworthiness could be addressed by having clinicians involved in study development [6].

[1] Corresponding Author: Melinda Wassell, email: drmelinda.wassell@gmail.com.

This study aims to explore the atemporal plausibility of clinician data in a structured EHR to determine fitness for use in predicting outcomes of work-related musculoskeletal disorders (WMSDs) using ML. Also, to engage clinicians to identify sources of inconsistency and suggest methods to improve EHR data quality.

2. Methods

The dataset was derived from a proprietary EHR used by chiropractors, physical therapists, and osteopaths at occupational health clinics across Australia. The dataset includes 57,570 patient complaints from 2014-2021. The organisation operates a value-based care model that provides reporting to employers and clinicians on injury trends. There is a high proportion of mandatory data fields, and the organization has internal governance processes to ensure data collection and reporting accuracy, as governance is an important parameter of data quality [7].

Following the method of the TRANSFoRm Zone Model [8], data was deidentified and extracted to a secure location for research.

Data quality must be assessed against the purpose of the data being used [9, 10], therefore a literature review of potential predictors of outcomes to WMSDs was conducted. The review formed the basis for determining relevant variables for analysis [11]. EHRs have yet to be widely used in prediction studies of WMSDs, which means there are potential variables captured in EHRs that are not studied. The organisation owning the dataset had previously completed a proof-of-concept ML model to predict the number of visits required to resolve WMSDs, and these variables were also considered.

The organisation works across various industries; however, the analysis was limited to a single industry (meat processing) to limit the natural variability of WMSDs across industries. Clinicians with less than 100 records in the dataset were excluded as they could be new, untrained clinicians. Dates for analysis were 1/1/2018-10/11/2021 due to EHR and organisational governance changes in 2017.

The clinicians had previously reported the EHR had a high documentation burden. Therefore, a team of senior clinicians were consulted to help inform the data analysis by understanding how they used the EHR. Consultation also engaged clinicians in understanding data quality, as focusing on EHR data quality has been shown to improve data quality [9, 12].

Missingness was assessed. Low missingness was expected for mandatory fields; however, just because clinicians must enter data doesn't mean the data will be plausible. For example, clinicians may document all patients' weight as 100kg.

Distribution comparisons [13] were conducted for each variable to verify atemporal plausibility. Continuous variables were grouped. Individual clinicians were compared to the clinician group to determine if there was correlation using the appropriate statistical test for the data type. Mean/median, standard deviations and outliers to 3 standard deviations were determined. Data quality findings were discussed with the clinical leadership team, and potential areas for improvement were categorised.

3. Results

Data processing yielded 15921 records. All variables were mandatory data capture excluding previous care, obstacles to recovery, smoking and exercise.

Table 1. Variables missingness and clinician data entry variability findings by percentage of records.

Variable	Missing	P values <.01	SD range	Alignment ^	Data quality improvement category suggestions
Age group	0	13/30	1.7-8.3	100	Not require
Gender	0	8/30	6.05	100	Update to gold standard
BMI group	47.8	19/30	1.7-16.1	100	Not required
Body Side	0	7/30	3.8-9.0	75	System rule limit values
Body Region	<0.1	5/30	3.1-6.2	78	Definitions, clinician training
Onset	<0.10	14/30	11.3-12	100	Not required
Numeric pain scales (3)	0	30/30	1.3-21.6	33.3	Standardize question, consider collection relevance
Symptom progress	<0.1	28/30	20.2-24.1	0	Clinician training
Mechanism of injury	0	28/30	2.8-20.4	60	Definitions, clinician training
Expected care visits	0	16/30	2.9-19.1	66.6	Clinician training
Work Related	0	16/30	9.3-9.6	100	Not required
Previous Care (free text)	99.1	4/15#	3.5-4.6	Unknown	Consider collection relevance, consider UI/UX
Obstacles to recovery	50.3	21/30	0.7-17.7	Unknown	Update to gold standard, clinician training
Smokes/day	75.9	12/30	1-11.7	100	Update to gold standard
Exercise (free text)	65.2	9/30	8.8-10.1	100	Update to gold standard

* % of individual clinician data entry statistically positively correlated to the clinician group data entry
^ % of values within each variable the clinical team found aligned with what was expected (local knowledge)
# Calculated based on clinicians who have entered data into variable and clinicians with sufficient records

Non-mandatory fields showed higher missingness, up to 99.1% for 'Previous Care' with only 44.4% of clinicians ever recording data in this field.

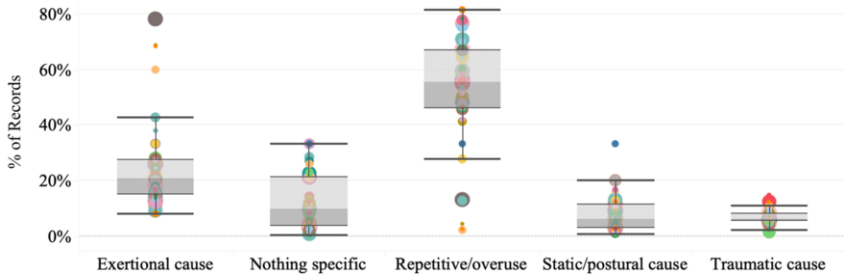

Figure 1. Clinician variability in use of Mechanism of injury variable. Circle size indicates number of records by clinician, colour indicates individual clinicians.

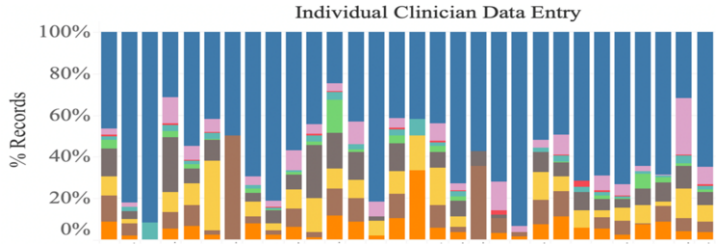

Figure 2. Clinician variability in use of Obstacles to recovery field. Color indicates category of obstacle. Columns are individual clinicians.

As seen from the Figure 1, the 'repetitive overuse' and 'exertional' values are dispersed. The clinical team recommended refining the definitions and training clinicians. Outlier records are 0.3% at 3 standard deviations. Statistically, this should not have much effect on modelling. However, there is a cluster within the outliers that an ML model may pick up on, which may lead to incorrect modelling. When findings were discussed with the clinical team, one clinician could identify that he was an outlier and had trained another clinician with similar outlier behaviour.

Analysis of 'obstacles to recovery' in Figure 2, showed inconsistencies in the way clinicians use the categories.

4. Discussion

Further analysis through patient audits are necessary to accurately determine missingness of non-mandatory fields. Previous internal clinician surveys indicated clinicians were more likely to enter data into the mandatory fields and skip non-mandatory fields to save time. This is evident in the various amounts of missing data between 'Obstacles to recovery' and 'Previous Care. 'Obstacles to recovery' is already utilized for reporting and training, with existing governance processes. 'Previous care' was not easily visible in the patient record once collected. Whilst it might be relevant to the proposed reuse, if the variable is not appropriate for primary EHR use, then consideration must be made of its relevance and effects on the clinician's time taken to enter data.

The assessment of 'Mechanism of Injury' showed high variability in correlation, which is likely responsible for the findings where most clinicians are statistically positively correlated to the clinician group. Reducing the inconsistency in the values could make it more likely that there is a negative association in outlying clinicians. It could be assumed that since there is a close correlation between clinicians in their use of 'traumatic' values, measures such as definitions and clinician training could improve the consistency of all values. Poor knowledge of how to use an EHR, such as unclear definitions of fields, is a known risk of quality issues in EHRs [3]. Fields with closer correlations, such as age groups, exhibit less consistent p-values among clinicians, highlighting a challenge in assessing correlation.

Clinicians know how they use the EHR. Rather than simply focusing on data quality from a technical perspective, clinicians were considered local knowledge experts and a significant part of the study design. The findings demonstrate that whilst statistically the data was plausible, clinicians could identify improvements that could be made to improve the capture and quality of data. It is proposed that data quality frameworks include clinicians in local knowledge assessment when assessing atemporal plausibility.

The study is part of a more extensive data quality analysis that assesses the intrinsic and extrinsic parameters [5, 7]. This study is limited to one method of determining atemporal plausibility. Analysis of variables using Delphi testing, data element agreement, validation checks, and validation to gold standards are required to thoroughly assess atemporal plausibility [13]. Whilst many of the findings of this study indicate many variables would be appropriate for reuse in predictive modelling, further study is needed.

5. Conclusions

The study highlights some inconsistencies in assessing the data quality parameter of atemporal plausibility that could lead to mistrust of the data by clinicians. Whilst statistical analysis found many variables were appropriate for reuse in predicting outcomes to WMSDs, clinicians found that there were inconsistencies in the data that affected plausibility. Engaging clinicians in data quality assessment can augment the assessment by identifying opportunities to improve quality.

Acknowledgment

The work was supported by Work Healthy Australia who provided the research dataset.

Ethics Approval

Consent and ethics approval was obtained CQUHREC Approval #0000023392.

References

[1] Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 2020;104:101822.
[2] Yoon A, Lee YY. Factors of trust in data reuse. Online information review. 2019;43(7):1245-62.
[3] Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018;20(5):e185.
[4] Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. EGEMS (Wash DC). 2017;5(1):14.
[5] 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.
[6] David W. Bates AA, Peter Schulam. Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence. Annals of Internal Medicine. 2020;172(11_Supplement): S137-S44.
[7] 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;28(7):1591-9.
[8] Kuchinke W, Ohmann C, Verheij RA, van Veen EB, Arvanitis TN, Taweel A, et al. A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. Int J Med Inform. 2014;83(12):941-57.
[9] van der Bij S, Khan N, Ten Veen P, de Bakker DH, Verheij RA. Improving the quality of EHR recording in primary care: a data quality feedback tool. J Am Med Inform Assoc. 2017;24(1):81-7.
[10] Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50 Suppl(0):S21-9.
[11] Thuraisingam S, Chondros P, Dowsey MM, Spelman T, Garies S, Choong PF, et al. Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment. BMC Medical Informatics and Decision Making. 2021;21(1).
[12] Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. Int J Med Inform. 2015;84(12):1094-8.
[13] 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;20(1):144-51.