Journal of Clinical Epidemiology 166 (2024) 111240

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Incomplete and possibly selective recording of signs, symptoms, and measurements in free text fields of primary care electronic health records of adults with lower respiratory tract infections

Merijn H. Rijk[a,*], Tamara N. Platteel[a], Marissa M.M. Mulder[a], Geert-Jan Geersing[a], Frans H. Rutten[a], Maarten van Smeden[b], Roderick P. Venekamp[a,1], Tuur M. Leeuwenberg[b,1]

[a]*Department of General Practice & Nursing Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

[b]*Department of Epidemiology & Health Economics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*

Accepted 5 December 2023; Published online 8 December 2023

Abstract

Objectives: To assess the completeness of recording of relevant signs, symptoms, and measurements in Dutch free text fields of primary care electronic health records (EHR) of adults with lower respiratory tract infections (LRTI).

Study Design and Setting: Retrospective cohort study embedded in a prediction modeling project using routine health care data of the Julius General Practitioners' Network of adult patients with LRTI. Free text fields of 1,000 primary care consultations of LRTI episodes between 2016 and 2019 were manually annotated to retrieve data on the recording of sixteen relevant signs, symptoms, and measurements.

Results: For 12/16 (75%) of the relevant signs, symptoms, and measurements, more than 50% of the values was not recorded. The patterns of recorded values indicated selective recording of positive or abnormal values. Recording rates varied across consultation type (physical consultation vs. home visit), diagnosis (acute bronchitis vs. pneumonia), antibiotic prescription issued (yes vs. no), and between practices.

Conclusion: In EHR of primary care LRTI patients, recording of signs, symptoms, and measurements in free text fields is incomplete and possibly selective. When using free text data in EHR-based research, careful consideration of its recording patterns and appropriate missing data handling techniques is therefore required.

Keywords: Electronic health record; Routine health care data; Natural language processing; Primary care; Lower respiratory tract infection; Missing data

1. Introduction

Electronic health records (EHR) hold information on large numbers of individual patients and contain structured data on health indicators such as medical history, coded disease episodes, presenting signs or symptoms, and treatments. Although EHR data are not primarily collected for research purposes, its potential for conducting research is increasingly recognized and the availability of longitudinal data on a large number of patients enables researchers to study relatively rare events [1,2]. Currently, the majority of EHR-based research is focused on structured (often limited) data, since manual extraction of information from unstructured free text fields is time-consuming and resource-consuming [2,3]. However, advances in natural language processing methods increasingly allow for automated retrieval of unstructured information from free text fields of the EHR, in which more detailed information on signs, symptoms, and measurements is often documented [2–4]. This increasing availability of unstructured data provides relevant opportunities for EHR-based research, for example on the prognosis of acute lower respiratory tract infections (LRTI) in primary care. Identifying LRTI patients at risk of hospitalization or mortality is challenging and existing prediction

Funding: This work was supported by ZonMw (grant number 08391052110003). ZonMw had no involvement in any part of the research process. AML was supported by NWO (grant number NWA.1418.22.008).

[1] Authors contributed equally to this work.

* Corresponding author. Department of General Practice & Nursing Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands. Tel.: +0031-88-75 681-81.

E-mail address: m.h.rijk@umcutrecht.nl (M.H. Rijk).

https://doi.org/10.1016/j.jclinepi.2023.111240

What is new?

Key findings

- In free text fields of primary care electronic health records (EHR), recording of signs, symptoms, and measurements is incomplete and possibly selective, potentially posing a challenge when used in EHR-based research.

What this adds to what was known?

- Although the potential of free text EHR data for retrieving research data is increasingly recognized, the quality of information captured in free text fields is rarely assessed in detail.
- This study assesses the completeness of recording of relevant signs, symptoms, and measurements in Dutch free text EHR data of primary care patients with a lower respiratory tract infection.

What is the implication and what should change now?

- When using free text primary care EHR data in research, careful consideration of recording patterns and appropriate missing data handling techniques is required.

models that can assist general practitioners (GP) in risk stratification—of which CRB-65 (confusion, respiratory rate, blood pressure and age $\geq$ 65 years) is the most promising—have not been properly validated (M. Rijk et al, unpublished data, 2023). Predictors of the CRB-65, as well as other potential relevant signs, symptoms, and measurements, are mainly recorded in free text fields [3]. However, prior to including free text variables in primary care EHR-based research, it is instrumental to gain insight in the type and quality of information captured in free text primary care EHR data and to assess the completeness of this information. Yet, this has been rarely done previously.

In this study, we therefore aimed to assess the completeness of recording of signs, symptoms, and measurements in free text fields of EHR of adult patients presenting to primary care with an LRTI. This provides insight on the recording in routine GP practice of such variables and informs on the quality and potential challenges in the use of free text data in primary care EHR-based research.

2. Study design and setting

2.1. Design and participants

For this study, we used primary care routine health care data from the Julius General Practitioners' Network (JGPN) [5]. The JGPN contains data on patient demographics, encoded diseases, and prescriptions (using the International Classification of Primary Care (ICPC) and Anatomical Therapeutic Chemical codes), coded measurements of vital parameters, laboratory results, and free text fields of consultations, and covers approximately 450,000 patients from both urban and rural general practices in the region of Utrecht, the Netherlands. This study is embedded in a project aimed at developing EHR-based prediction models for adverse outcomes in adult patients presenting with LRTI to primary care, which has been reported on elsewhere [6]. In short, the cohort comprises of patients aged 40 years and older consulting their GP with an LRTI episode (defined as ICPC registration of either acute bronchitis [R78] or pneumonia [R81]) between 1 January 2016 and 31 December 2019. An episode starts at the day of the first LRTI-related consultation (i.e., index consultation), and consecutive episodes within individual patients are separated by a period of at least 28 days without LRTI-related consultations. For this study, a nonstratified random sample of 1,000 episodes from our JGPN cohort was drawn. In case of multiple episodes per individual patient, only the first episode was included and we only included episodes of which the index consultation was a face-to-face appointment, since these are likely to contain information on physical examination and vital and laboratory measurements.

2.2. Data collection

For all 1,000 episodes, information on patient demographics (age, sex), comorbidities (pulmonary diseases, diabetes, heart failure, and history of malignancy or cerebrovascular accident), smoking status and medication use (immunosuppressants, inhalation medication) was collected at the day of index consultation. Smoking status was automatically extracted using a free text algorithm that has been developed within JGPN [7]. Using free text fields and structured input on measurements, we extracted data on signs, symptoms, and measurements recorded at index consultations which may be of value for predicting poor outcome in primary care LRTI patients. These include patient-reported symptoms (cough, shortness of breath, sputum, chest pain, chills, fever, and confusion), GP-reported signs (ill appearance, confusion, crackles on lung auscultation), and vital and laboratory measurements (blood pressure, body temperature, respiratory rate, heart rate, oxygen saturation, and C-reactive protein). Recording of these variables was manually retrieved from free text fields of index consultations by one author (AM). During this process, a second author (MR) manually cross-checked the first 100 annotated episodes (10% of the total sample) to assure correct coding of free text derived variables in the remaining sample, with a high inter-rater reliability (Cohen's Kappa 0.98 [95% confidence interval (CI) 0.97—0.99], indicating almost perfect agreement). When recorded, the absence or presence of signs and symptoms (e.g., cough yes/no)

and the values of measurements were documented. Antibiotic prescription at index consultation was extracted using the Anatomical Therapeutic Chemical codes of LRTI-related prescriptions including amoxicillin, amoxicillin/clavulanate, azithromycin, clarithromycin, doxycycline, and erythromycin [8].

2.3. Statistical analysis

We performed an exploratory analysis of the extracted data on signs, symptoms, and measurements, in which the proportion of recording, the distribution of positive and negative values of recorded signs and symptoms, and the distribution of recorded measurements were descriptively summarized. We explored differences in the recording of signs, symptoms, and measurements across the strata of consultation type, diagnosis, antibiotic treatment at index consultation, and age, and we assessed heterogeneity in recording patterns between general practices.

Proportions of registered signs, symptoms, and measurements were compared using a two-sample Z-test, and mean number of recordings were compared using the Wilcoxon rank-sum test. Correlation between continuous variables was assessed using Pearson's correlation coefficient. All analyses were performed in R version 4.2.2 [9].

3. Results

3.1. Characteristics of cohort

Characteristics of the cohort are summarized in Table 1. The mean age of the patients was 63.0 (standard deviation (SD) 13.8) years, and 56.0% were female. Most patients (56.3%) were diagnosed with pneumonia and 31.9% were treated with antibiotics at index consultation.

3.2. Recording of signs, symptoms, and measurements

On average, 3.6 signs and symptoms (median: 4) were recorded per patient (Fig. 1). Lung auscultation (86.1%) and cough (76.6%) were recorded most frequently whereas information on chills (4.4%) and confusion (3.5%) was recorded least frequently (Table 2). When recorded, patient reported symptoms were mostly present—most notably cough (98.4%), chills (95.5%), sputum (91.1%) and chest pain (78.3%)—indicating selective recording.

An average number of 1.9 measurements (median: 2) was recorded per patient (Fig. 1). For all measurements, more than 50% of the values were not recorded (Table 2). Body temperature (45.4%), oxygen saturation (45.1%), and heart rate (40.0%) were recorded most frequently whereas respiratory rate (5.7%) was infrequently recorded. The distributions of recorded measurement values—especially respiratory rate, oxygen saturation, and CRP—reveal a substantial number of abnormal values, potentially indicating selective recording in those more severely affected (Fig. 2).

Table 1. Baseline characteristics of cohort

Characteristic	Total = 1,000
Mean age (SD)	63 (13.8)
Sex - female	560 (56.0%)
Year	
2016	343 (34.3%)
2017	241 (24.1%)
2018	246 (24.6%)
2019	170 (17.0%)
Consultation type	
Physical consultation	850 (85.0%)
Home visit	150 (15.0%)
Diagnosis	
Acute bronchitis	437 (43.7%)
Pneumonia	563 (56.3%)
Smoking status	
Current	266 (26.6%)
Former	209 (20.9%)
Never	202 (20.2%)
No information	323 (32.3%)
Comorbidities	
COPD	102 (10.2%)
Asthma	121 (12.1%)
Malignancy[a]	78 (7.8%)
Diabetes	52 (5.2%)
Heart failure	37 (3.7%)
CVA	38 (3.8%)
Medication use	
Inhalation medication [b]	45 (4.5%)
Immunosuppressants or systemic corticosteroids	9 (0.9%)
Antibiotic prescription (at index consultation)[c]	319 (31.9%)

Abbreviations: COPD, chronic obstructive pulmonary disease; CVA, cerebrovascular accident; SD, standard deviation.

[a] Excluding malignancies of the skin.

[b] Includes long acting bronchodilators, corticosteroid inhalers, and combinations.

[c] Includes only antibiotic prescriptions that are appropriate for treating lower respiratory tract infections based on primary care guidelines, that is, amoxicillin, amoxicillin/clavulanate, azithromycin, clarithromycin, doxycycline, and erythromycin.

3.3. Recording patterns across subgroups

The mean number of recorded signs and symptoms was significantly higher in patients with ICPC registration of pneumonia compared to acute bronchitis (3.7 vs. 3.5, $P = 0.006$) and in those receiving immediate antibiotic treatment compared to those who did not (3.9 vs. 3.4, $P < 0.001$) (Table 3). On average, the number of recorded measurements was significantly higher at home visits

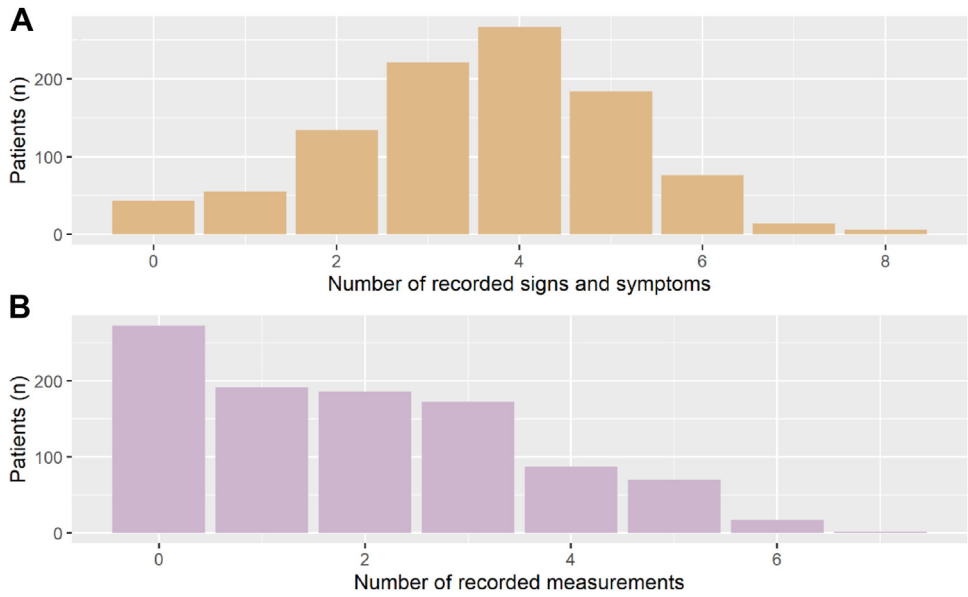

Fig. 1. Number of (A) signs and symptoms and (B) measurements recorded per individual patient. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

compared to physical consultations (2.7 vs. 1.8, $P < 0.001$), in patients with pneumonia compared to acute bronchitis (2.2 vs. 1.5, $P < 0.001$), and in those receiving immediate antibiotic treatment compared to those who did not (2.2 vs. 1.8, $P < 0.001$). There was a significant difference in the recording of 9/15 variables between

Table 2. Recording and distribution of signs, symptoms, and measurements in free text fields of electronic medical records

Signs and symptoms

	Recording (%)			
Variable	**Pneumonia**	**Acute bronchitis**	**Overall**	**Positive if recorded (%)**
Patient reported				
Cough	71.9	82.6	76.6	98.4
Fever	57.0	50.3	54.1	57.7
Shortness of breath	54.4	51.7	53.2	70.9
Sputum	26.1	30.7	28.1	91.1
Chest pain	22.9	13.5	18.8	78.7
Chills	5.7	2.7	4.4	95.5
Patient or GP reported				
Confusion	5.3	1.1	3.5	17.1
GP reported				
Crackles (auscultation)	83.7	89.2	86.1	26.6
Ill appearance	39.1	24.7	32.8	39.3

Measurements

	Recording (%)			
Variable	**Pneumonia**	**Acute bronchitis**	**Overall**	**Median value (IQR)**
Body temperature	52.0	36.8	45.4	37.2 (36.8–38.0) °C
Oxygen saturation	51.0	37.5	45.1	97 (95–98)%
Heart rate	46.9	31.1	40.0	88 (76–100) beats/min
CRP	20.1	16.5	18.5	39 (10–94) ml/mL
Systolic BP	21.7	12.1	17.5	132 (120–150) mmHg
Diastolic BP	21.7	12.1	17.5	80 (70–84.5) mmHg
Respiratory rate	6.4	4.8	5.7	18 (16–20) breaths/min

Abbreviations: BP, blood pressure; CRP, C-reactive protein; GP, general practitioner.

Table 3. Differences in mean number of recordings between LRTI episodes stratified according to consultation type, diagnosis, and immediate antibiotic prescription

Consultation type			
	Mean number of recordings		
Variable type	**Home visit**	**Physical consultation**	***P* value**[a]
Signs and symptoms	3.7	3.5	0.106
Measurements	2.7	1.8	<0.001
Diagnosis			
	Mean number of recordings		
Variable type	**Pneumonia**	**Acute bronchitis**	***P* value**[a]
Signs and symptoms	3.7	3.5	0.006
Measurements	2.2	1.5	<0.001
Treatment			
	Mean number of recordings		
Variable type	**Immediate AB**	**No immediate AB**	***P* value**[a]
Signs and symptoms	3.9	3.4	<0.001
Measurements	2.2	1.8	<0.001

Abbreviations: AB, antibiotics; LRTI, lower respiratory tract infection.
[a] Based on Wilcoxon rank sum test.

consultation type (home visit vs. physical consultation), 12/15 variables between diagnosis (pneumonia vs. acute bronchitis), and 8/15 variables between antibiotic prescription issued at index consultation (yes vs. no) (Table S1). Shortness of breath was the only variable that did not significantly differ across all three strata. Age was weakly negatively associated with the number of registered signs and symptoms included in our study (Pearson's correlation coefficient (r) −0.07, 95% CI −0.13 to −0.01) but showed no significant association with the number of registered

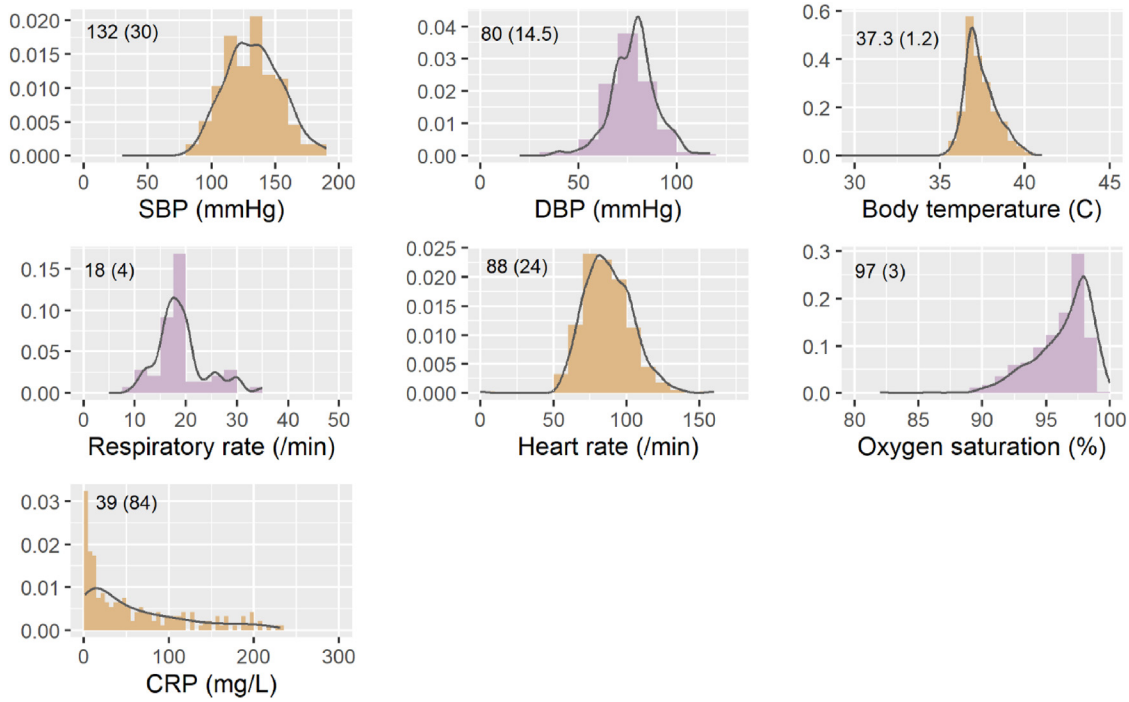

Fig. 2. Distribution of recorded measurements with density curve and median (interquartile range). Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; C, Celsius; min, minute; CRP, C-reactive protein. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

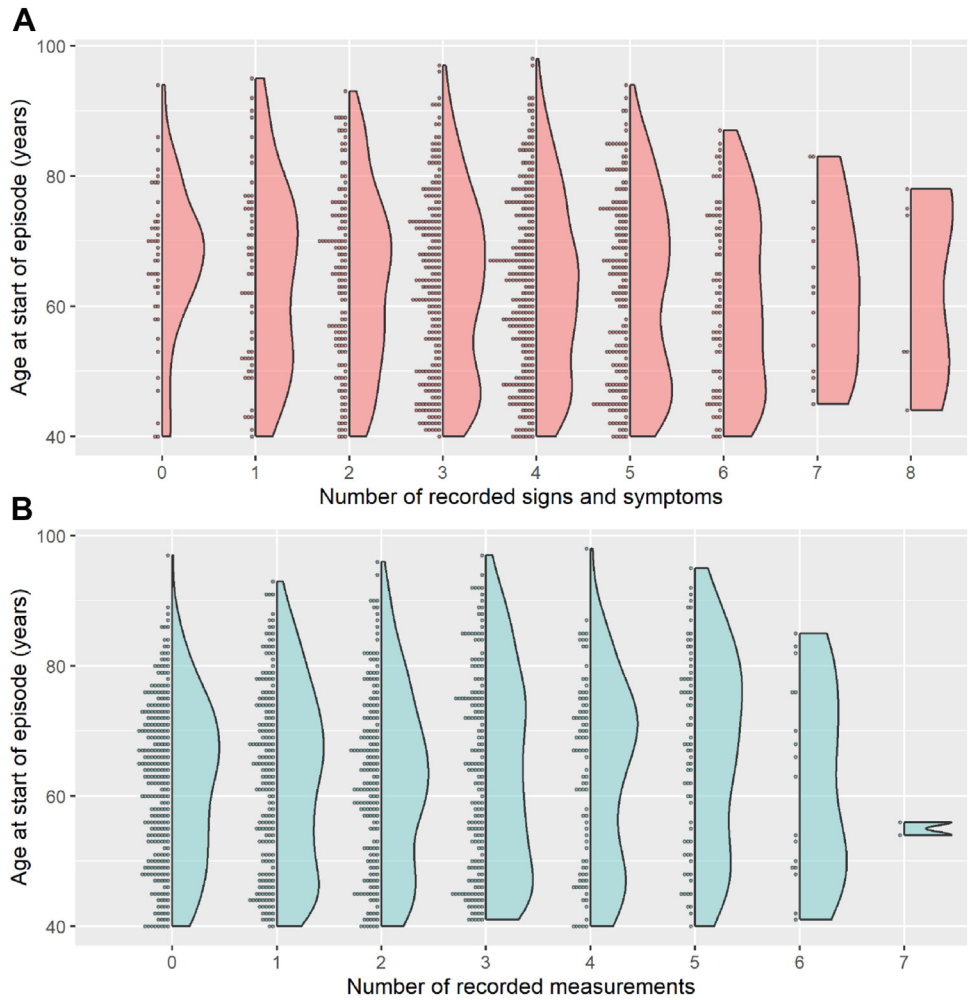

Fig. 3. Association between age and (A) number of signs and symptoms, and (B) number of measurements recorded. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

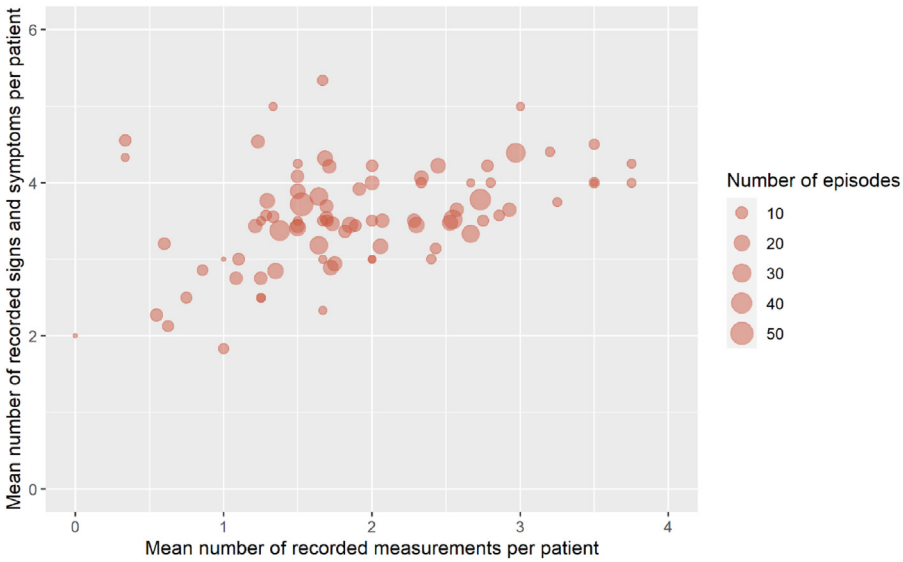

Fig. 4. Association between mean number of signs and symptoms and mean number of measurements recorded, on a general practice level. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

measurements (r 0.06, 95% CI 0.00−0.12) (Fig. 3). The number of recorded signs and symptoms was positively associated with the number of recorded measurements (r 0.30, 95% CI 0.24−0.35).

Summarizing the recording patterns of different primary care practices, a mean number of 3.6 (SD 0.7, range 1.8−5.3) signs and symptoms and 1.9 (SD 0.8, range 0.0−3.8) measurements were recorded per patient per practice. On a primary care practice level, the mean number of recorded signs and symptoms per patient was positively associated with the mean number of recorded measurements per patient on a primary care practice level (r 0.44, 95% CI 0.25−0.60) (Fig. 4).

4. Discussion

4.1. Main findings

In this study, we aimed to assess the completeness of recording of relevant signs, symptoms, and measurements in free text EHR data of primary care LRTI patients. Our most notable finding is the substantial amount of nonrecorded values. Apart from incomplete recording, our study reveals several issues indicating selective recording of these variables, potentially posing a challenge when using these data in EHR-based research. First, the distribution of recorded values suggests that signs, symptoms, and measurements are more likely to be recorded in case of positive or abnormal findings. Second, we found that their recording is associated with observed factors such as consultation type, diagnosis, and antibiotic treatment, raising uncertainty on potential associations with nonobserved factors. Lastly, we observed heterogeneity in recording behavior between primary care practices.

4.2. Comparison with existing literature

Primary care guidelines currently aid GPs in identifying LRTI patients with increased risk of adverse outcomes by highlighting factors associated with such outcomes, including several signs, symptoms, and measurements. for example, the Dutch primary care guideline includes ill appearance, confusion, lung auscultation, respiratory rate, heart rate, and blood pressure in its severity assessment [8]. UK guidance on pneumonia patients issued by the National Institute of Health and Care Excellence suggests to stratify risk of poor prognosis in primary care LRTI patients based on CRB-65 [10]. Nevertheless, the findings in our study indicate that the factors addressed in these guidelines are not routinely recorded in LRTI patients, which is largely in line with findings of previous studies. A prospective evaluation of the potential use of CRB-65 in primary care LRTI patients conducted in 13 European countries reported that respiratory rate (22.7%) and blood pressure (31.9%) were infrequently measured as part of routine practice, whereas confusion (99.8%) was recorded in almost all patients as it was part of a standard case report form [11]. Another retrospective UK EHR-based study among primary care patients with community-acquired pneumonia found that complete recording of CRB-65 at index consultation was 0.4%; confusion (0.2%), respiratory rate (3.6%), and blood pressure (17.6%) were all infrequently recorded [12]. Of note, in a primary care-based prospective cohort study of patients with LRTI with near-complete recording, the positivity rates of signs and symptoms were substantially lower than in our study (e.g., fever and chest pain reported in 38.1% and 37.0% vs. 57.7% and 78.7%, respectively) and the distributions of measurements revealed less abnormal values [13]. Hence, selective recording of signs, symptoms, and measurements in our study is likely. This might be explained by the fact that GPs both use their severity assessment to selectively assess these factors in a subset of patients and mainly report positive or abnormal values. Extrapolating these positivity rates to our data does, however, not imply that all nonrecorded values in our study are negative or normal.

When considering free text variables for use in observational EHR-based prognostic research, nonrecorded values could be considered 'missing.' Proper handling of missing data and careful consideration of its mechanisms is therefore essential [14]. Our study identifies two issues that might pose a challenge when using free text EHR data for this purpose: (i) the substantial amount of nonrecorded values and (ii) the selective recording of variables. Although the maximum proportion of missing data on variables that is generally accepted in prognostic studies has been subject of debate, some of the observed proportions of missingness in our study are likely to pose a challenge when applying techniques such as multiple imputation (MI) [15−17]. Selective recording could potentially be even more problematic, since this might violate assumptions regarding the type of missingness that should be met before applying MI (Box 1) [18]. In our study, we identified several variables (i.e., consultation type, diagnosis, and immediate antibiotic treatment) that are associated with the proportion of missingness−indicating missing at random −and it is likely that even more observed variables

Box 1 Different types of missing data

Missing data are commonly classified into three distinct types [18].

- Missing completely at random (i.e., no systematic differences between observed and missing values)
- Missing at random (i.e., systematic differences between observed and missing values which can be explained by other variables in the observed data)
- Missing not at random (i.e., systematic differences between observed and missing values remain after taking other variables in the observed data into account).

are associated with the level of missingness. e.g., our observation that signs, symptoms, and measurements are presumably recorded more often in case of positive or abnormal findings might indicate missing not at random , which violates one of the assumptions of MI [18]. Alternative methods for handling incomplete recording of variables could be by considering missing values as 'negative' or 'normal' (i.e., zero imputation) or to include a missing indicator [19,20]. However, when such variables are eventually included in a prediction model their assessment becomes routine care, and GPs will increasingly assess (and record) negative or normal values inherently altering the mechanism of missingness. This may lead to changing predictor effects, which is also referred to as 'the curse of knowing' [21].

4.3. Strengths and limitations

Strengths of our study include the large sample that was derived from a cohort representative of Dutch adult primary care LRTI patients, the selection of relevant signs, symptoms, and measurements that was based on a recent systematic review, and the manual annotation of the variable values from free text EHR data by two authors with high inter-rater reliability.

Some important limitations should, however, be recognized. First, our study should foremost be regarded as an exploratory analysis of the completeness of recording in free text primary care EHR data, and it was not our aim to draw strong conclusions regarding the value of free text EHR data for future research. Nevertheless, we believe our results to be relevant as they highlight some important issues to consider when using free text EHR data in research. Second, it is important to emphasize that the absence of recorded signs, symptoms, and measurements does not necessarily imply that these were not assessed by the GP. It is likely that GPs do not routinely register all information that has been acquired during a consultation while our findings are solely based on recorded data. Therefore, no firm conclusions should be drawn regarding the patient characteristics that GPs routinely assess in primary care LRTI patients. Third, our findings only reflect routine primary health care for LRTI patients during regular working hours, since we did not include out-of-hours consultations. However, even in acutely ill patients presenting to out-of-hours primary care facilities vital parameters such as respiratory rate are often not measured according to a previous report [22]. Lastly, our data do not include information on the course of the LRTI episodes and whether adverse outcomes such as hospitalization or mortality occurred. As these are objective indicators of disease severity, it would have been valuable to assess the association between the recording of signs, symptoms, and measurements and adverse outcomes.

4.4. Implications for future research

In spite of the descriptive nature of this study, our findings are relevant to other future EHR-based research aiming to include free text variables. Before applying natural language processing to free text EHR data, it is important to gain insight into the completeness of recording of free text variables and identify potential challenges that come with its use. Our findings imply that, prior to including free text variables, recording patterns should be carefully assessed and appropriate missing data handling techniques should be considered based on these patterns. To place our findings in the context of future (prognostic) research and highlight the challenges of using free text EHR data for validating existing prediction models, we have used the CRB-65 model as an illustrative example in Box 2 and Fig. 5. Nevertheless, the potential value of information captured in free text fields justifies further exploration [23]. Therefore, to properly assess the added value of free text variables in prognostic research, future studies should compare the performance of a prediction model developed with only structured data to a model that also includes free text predictors using appropriate methods [24], and with proper handling of missing data depending on recording patterns and how the model will be applied in practice.

CRediT authorship contribution statement

M.H. Rijk: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. **T.N. Platteel:** Conceptualization, Methodology, Supervision, Writing – review & editing. **A.M.M. Mulder:** Conceptualization, Data curation, Formal analysis, Writing – review & editing. **G.J. Geersing:** Conceptualization, Methodology, Writing – review & editing. **F.H. Rutten:** Conceptualization, Methodology, Writing – review & editing. **M. van Smeden:** Conceptualization, Methodology, Writing – review & editing. **R.P. Venekamp:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **A.M. Leeuwenberg:**

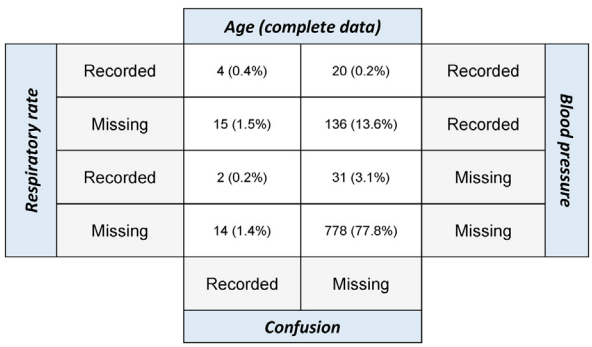

Fig. 5. Recording of CRB-65 predictors on a patient level based on manual annotation of 1,000 primary care LRTI consultations.

Box 2 Recording of CRB-65 predictors: an illustrative example

To place our findings in the context of future (prognostic) research and highlight the challenges of using free text EHR data for validating existing prediction models, we summarized the recording of CRB-65 predictors (a model developed to predict mortality in pneumonia patients) on a patient level. In our sample, there was no missing data on age, but confusion (3.5%), respiratory rate (5.7%), and blood pressure (17.5%) were all incompletely recorded. Overall, 778 (77.8%) LRTI patients only had information on age available whereas 4 (0.4%) had complete recording of CRB-65 items. Due to the substantial amount of nonrecorded values (i.e., equivalent to missing data since this information is not available elsewhere), external validation of CRB-65 using EHR-data is challenging and would require tailored missing data handling techniques based on the observed recording patterns and underlying assumptions. For example, one could argue to consider nonrecorded values of confusion as 'negative' or 'normal' [20]. Whether this assumption also holds for nonrecorded values of respiratory rate and blood pressure might be less certain.

Conceptualization, Funding acquisition, Methodology, Supervision, Writing − review & editing.

Data availability

Data access is restricted, and has been granted under license of the current study. Data are therefore only available from the authors upon reasonable request and after formal permission of the JGPN.

Declaration of competing interest

None declared by all authors.

Acknowledgments

The authors thank Marloes *M. van* Beurden for extracting the required data from the JGPN database.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.111240.

References

[1] Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61−81.
[2] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395−405.
[3] Fu S, Wang L, Moon S, Zong N, He H, Pejaver V, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: a scoping review. Clin Transl Sci 2022;16:398−411.
[4] Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc 2022;29:1292−302.
[5] Smeets H, Kortekaas M, Rutten F, Bots M, van der Kraan W, Daggelders G, et al. Routine primary care data for scientific research, quality of care programs and educational purposes: the Julius General Practitioners' Network (JGPN). BMC Health Serv Res 2018;18:735.
[6] Rijk MH, Platteel TN, Geersing GJ, Hollander M, Dalmolen BLGP, Little P, et al. Predicting adverse outcomes in adults with a community-acquired lower respiratory tract infection: a protocol for the development and validation of two prediction models for (i) all-cause hospitalisation and mortality and (ii) cardiovascular outcomes. Diagn Progn Res 2023;7:23.
[7] de Boer AR, de Groot MCH, Groenhof TKJ, van Doorn S, Vaartjes I, Bots ML, et al. Data mining to retrieve smoking status from electronic health records in general practice. Eur Heart J Digit Health 2022;3(3):437−44.
[8] Verheij T, Hopstaken R, Prins J, Salomé P, Bindels P, Ponsioen B, et al. Nederlands Huisartsen Genootschap. 2020. NHG-Standaard Acuut hoesten. Available at https://richtlijnen.nhg.org/standaarden/acuut-hoesten#volledige-tekst. Accessed May 25, 2022.
[9] R Core Team. R: a language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available at https://www.R-project.org/. Accessed June 2, 2023.
[10] National Institute for Health and Care Excellence (NICE). Pneumonia in adults: diagnosis and management (Clinical guideline 191) [Internet] 2014. Available at www.nice.org.uk/guidance/cg191. Accessed June 2, 2023.
[11] Francis NA, Cals JW, Butler CC, Hood K, Verheij T, Little P, et al. Severity assessment for lower respiratory tract infections: potential use and validity of the CRB-65 in primary care. Prim Care Respir J 2012;21(1):65−70.
[12] Launders N, Ryan D, Winchester C, Skinner D, Konduru PR, Price DB. Management of community-acquired pneumonia: an observational study in UK primary care. Pragmat Obs Res 2019;10:53−65.
[13] Moore M, Stuart B, Lown M, Van Den Bruel A, Smith S, Knox K, et al. Predictors of adverse outcomes in uncomplicated lower respiratory tract infections. Ann Fam Med 2019;17(3):231−8.
[14] Nijman SWJ, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs JJL, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol 2022;142:218−29.
[15] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Med Res Methodol 2017;17:162.
[16] Lee KJ, Carlin JB. Recovery of information from multiple imputation: a simulation study. Emerg Themes Epidemiol 2012;9:3.
[17] Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol 2019;110:63−73.

[18] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;339:157−60.
[19] Sperrin M, Martin GP. Multiple imputation with missing indicators as proxies for unmeasured variables: simulation study. BMC Med Res Methodol 2020;20:185.
[20] Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013;1(3):7.
[21] Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. Diagn Progn Res 2020;4(1):8.
[22] Loots FJ, Dekker I, Wang RC, van Zanten ARH, Hopstaken RM, Verheij TJM, et al. The accuracy and feasibility of respiratory rate measurements in acutely ill adult patients by GPs: a mixed-methods study. BJGP Open 2022;6(4): BJGPO.2022.0029.
[23] Seinen TM, Kors JA, van Mulligen EM, Fridgeirsson E, Rijnbeek PR. The added value of text from Dutch general practitioner notes in predictive modeling. J Am Med Inform Assoc 2023;30:1973−84.
[24] Cook NR. Quantifying the added value of new biomarkers: how and how not. Diagn Progn Res 2018;2(1):14.