Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study

Dijk, W.B. van; Fiolet, A.T.L.; Schuit, E.; Sammani, A.; Groenhof, T.K.J.; Graaf, R. van der; ... ; Mosterd, A.

Journal of Clinical Epidemiology 132 (2021) 97−105

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study

Wouter B. van Dijk[a,*,1], Aernoud T.L. Fiolet[b,c,1], Ewoud Schuit[a], Arjan Sammani[c], T. Katrien J. Groenhof[a], Rieke van der Graaf[d], Martine C. de Vries[e], Marco Alings[f,g], Jeroen Schaap[f,g], Folkert W. Asselbergs[c,h,i], Diederick E. Grobbee[a], Rolf H.H. Groenwold[j], Arend Mosterd[a,b,g]

[a]*Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands*

[b]*Department of Cardiology, Meander Medical Center, Amersfoort, the Netherlands*

[c]*Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands*

[d]*Department of Medical Humanities, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands*

[e]*Department of Medical Ethics and Health Law, Leiden University Medical Center, Leiden University, Leiden, the Netherlands*

[f]*Department of Cardiology, Amphia Hospital, Breda, the Netherlands*

[g]*Dutch Network for Cardiovascular Research (WCN), Utrecht, the Netherlands*

[h]*Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom*

[i]*Health Data Research UK and Institute of Health Informatics, University College London, London, United Kingdom*

[j]*Department of Clinical Epidemiology, Leiden University Medical Center, Leiden University, Leiden, the Netherlands*

Accepted 18 November 2020; Published online 25 November 2020

Abstract

Objective: This study aimed to validate trial patient eligibility screening and baseline data collection using text-mining in electronic healthcare records (EHRs), comparing the results to those of an international trial.

Study Design and Setting: In three medical centers with different EHR vendors, EHR-based text-mining was used to automatically screen patients for trial eligibility and extract baseline data on nineteen characteristics. First, the yield of screening with automated EHR text-mining search was compared with manual screening by research personnel. Second, the accuracy of extracted baseline data by EHR text mining was compared to manual data entry by research personnel.

Results: Of the 92,466 patients visiting the out-patient cardiology departments, 568 (0.6%) were enrolled in the trial during its recruitment period using manual screening methods. Automated EHR data screening of all patients showed that the number of patients needed to screen could be reduced by 73,863 (79.9%). The remaining 18,603 (20.1%) contained 458 of the actual participants (82.4% of participants).

In trial participants, automated EHR text-mining missed a median of 2.8% (Interquartile range [IQR] across all variables 0.4−8.5%) of all data points compared to manually collected data. The overall accuracy of automatically extracted data was 88.0% (IQR 84.7−92.8%).

Funding: This work was supported by the Netherlands Organisation for Health Research and Development (ZonMW) (grant number 91217027). A. Sammani was funded by the University Medical Center Utrecht Alexandre Suerman Stipendium. Folkert Asselbergs was supported by UCL Hospitals NIHR Biomedical Research.

Conflicts of interest: Rieke van der Graaf reported being a member of an independent ethical advisory committee to Sanofi. All other authors did not report any conflicts of interest.

Author statement: 1) Conceived and designed the experiments: van Dijk, Fiolet, Schuit, Grobbee, Groenwold, Mosterd. 2) Performed the experiments; van Dijk, Fiolet, Sammani, Groenhof. 3) Analyzed and interpreted the data; van Dijk, Fiolet, Schuit. 4) Contributed reagents, materials, analysis tools or data: van der Graaf, de Vries, Alings, Schaap, Asselbergs. 5) Wrote the paper: van Dijk, Fiolet.

[1] Shared first authorship.

* Corresponding author. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, the Netherlands. Tel.: +31 (0)6 12 43 45 58.

E-mail address: W.B.vanDijk-7@umcutrecht.nl (W.B. van Dijk).

https://doi.org/10.1016/j.jclinepi.2020.11.014

Conclusion: Automatically extracting data from EHRs using text-mining can be used to identify trial participants and to collect baseline information.

Keywords: Text-mining; Data-mining; Electronic healthcare records (EHRs); Electronic medical records (EMRs); Cardiovascular; Trials; Multicenter; Recruitment; Screening; Data-collections; LoDoCo2

1. Introduction

Clinical research requires highly detailed information on large numbers of subjects, often acquired by many investigators and supporting staff. In particular, prospective research such as registries and randomized clinical trials (RCT) need to comply with high standards of data validity [1,2]. Scientific and regulatory requirements make such endeavors laborious and increase costs to a level only large companies are able to meet.

Cardiovascular outcome trials with moderate to low absolute risks nowadays require over 10,000 participants and are estimated to cost between 35,000 and 45,000 US dollars per participant, with total costs up to half a billion US dollars for conduct [3,4]. A major part of these costs is attributable to participant recruitment and follow-up, for a large part comprising data collection [5,6]. Standing practice for clinical trials is that dedicated personnel enters source data in distinct (electronic) clinical report forms (CRFs). This data, however, is generally already collected in clinical care and available in electronic healthcare records (EHRs), thus creating overlapping copies of data that are already available (Fig. 1A).

Automated EHR data-mining may provide a valuable method to complement or even substitute current data collection methods [7], which could save up to one-third of recruitment costs [8]. In recent years, several supervised patient-diagnosis registries with labeled clinical data emerged to improve trial efficiency [9]. The use of automatically collected EHR data in trials, however, is still very limited [10]. Conventional data collection methods generally involve retrieving information through researcher-patient interviews and manual data extraction. After retrieval, data is then entered manually in electronic data capture (EDC) systems as part of CRFs. Data quality is guaranteed up to a certain level by automated control processes and internal and external monitoring [11]. If EHR data are to be used to identify participants or as an alternative data source, these data should be of sufficient quality. High data quality is paramount, yet will differ per objective. The accuracy level is relative to the nature of the data. Outcome data that is used to estimate a treatment effect requires higher fidelity than baseline data [12].

We hypothesized that patients eligible for trial participation can be effectively identified on information already present in EHRs using automated text-mining. Second, we hypothesized that the majority of data collected for the purpose of the trial is also already available in EHRs. If extracted automatically with acceptable accuracy, the extensive manual entry by investigators in EDCs could be reduced. If true, data collection efforts could focus on information not available from EHRs and reduce manual EHR-to-EDC data duplication that is now common (Fig. 1).

2. Methods

This study was a multicenter, multi-EHR-vendor validation study to assess the accuracy of automated EHR text-mining for trial participant screening and baseline data collection. As a reference standard, we used manual participant screening and data collection by manual data entry in EDCs, which is the current standard for most RCTs.

First, all patients who visited the outpatient cardiology clinics of the three participating medical centers during the recruitment phase (October 1, 2016, to December 1, 2018) of the LoDoCo2 trial were automatically and anonymously screened retrospectively for eligibility of participation in the trial according to its inclusion and exclusion criteria (Fig. 2). The yield of eligible patients via this

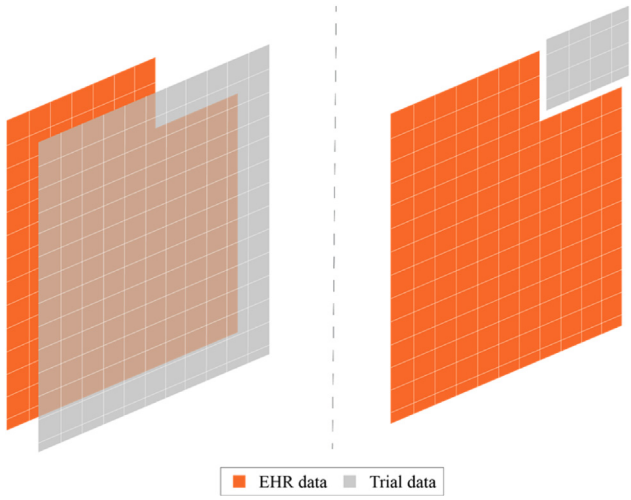


Fig. 1. Layers of data collected during trials (left: required trial data collection when not using EHR data in perspective to data available in EHR; right: (theoretical) required trial data collection when using EHR data).

Key findings

- Compared to conventional methods, automated text-mining in electronic healthcare records (EHRs) can substantially reduce the number of patients that need to be screened for trial enrollment and the amount of labor to collect data.

What this adds to what was known

- Previous studies mining parts of EHRs showed mixed results for text-mining methods and mainly focussed on observational and registry data-collection.
- This study shows that integral text-mining of EHRs yields good results for trial participant screening and data-collection.

What is the implication and what should change now

- Clinical trials should consider automated text-mining methods to supplement current participant screening and data-collection methods.

method was compared to those actually included in the trial by manual screening for trial participation. Second, baseline characteristics were automatically collected for all trial participants, and accuracy was assessed against manually collected data.

2.1. The LoDoCo2 trial

Conventional participant identification and data collection methods used in the international clinical trial LoDoCo2 were used as the reference standard. The LoDoCo2 trial was chosen as it represents a prototype large international multicenter cardiovascular outcome trial.

In short, the LoDoCo2 trial was a randomized, investigator-initiated international, multicenter study that investigated whether colchicine 0.5 mg once daily as compared to placebo in patients with stable coronary artery disease reduces the incidence of major adverse cardiovascular events [13]. The trial's recruitment started in December 2016 and was completed in December 2018. The trial methodology and results have been reported before [14].

2.2. Study population

This study was based on the data of patients visiting the cardiology outpatient clinics of three large Dutch medical centers. The medical centers were selected to represent the major EHR software vendors in the Netherlands (Epic [Hospital A], ChipSoft [Hospital B], CSC Care solutions [Hospital C]; cumulatively used in 80% of the Dutch hospitals and almost 10% of the hospitals worldwide [15,16]).

Participants of the LoDoCo2 trial were retrieved on their trial identification number and unique on-site identifier as recorded in their EHR files. Participants for which no trial identifiers were reported in the EHR were ignored since they could not be linked to CRF data functioning as the reference standard.

2.3. Participant identification methods

2.3.1. Automatic, using text-mining from EHRs

A Boolean retrieval query to obtain the required data was developed in adherence with the eligibility criteria of the LoDoCo2 trial by two authors (WBvD and ATLF) (Supplement 1a). For developing the query a graphic user interface data mining tool with text-mining features was used (CTcue, version 2.0.12; Amsterdam, The Netherlands). This data mining tool integrally searched structured and unstructured EHR data (including clinical

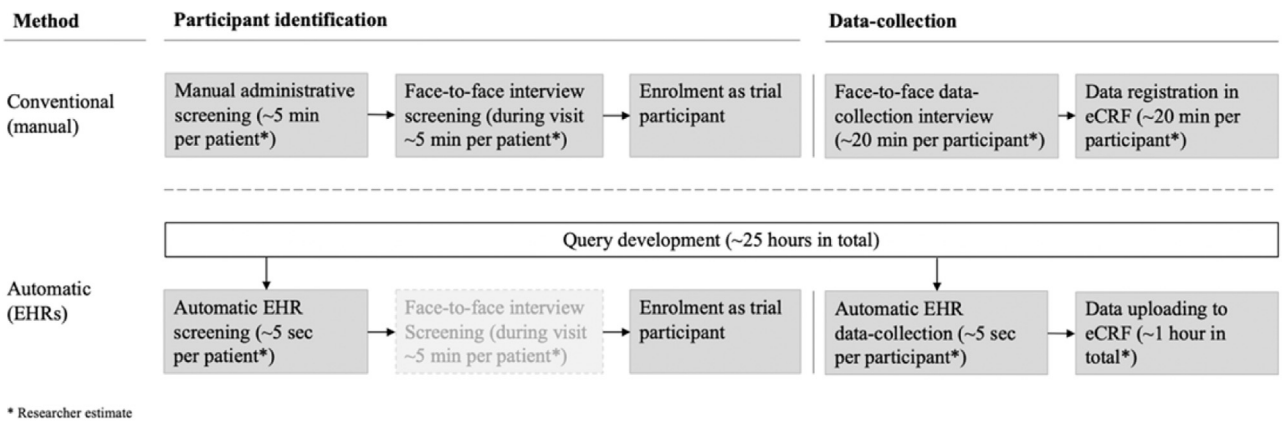


Fig. 2. Overview of the process of conventional and automated participant identification and data collection and the associated estimated time of these processes.

letters, in-hospital consultations, procedures, diagnostic tests, and drug prescriptions).

Both authors who developed the query were considered to have content expertise from their medical backgrounds and had extensive experience in query development. Additionally, one of these authors (ATLF) was also a lead investigator of the LoDoCo2 trial.

The query consisted of regular expressions of the eligibility criteria as given by the LoDoCo2 trial, their synonyms, and negations (e.g., ''no hypertension'' instead of ''hypertension''). Synonyms were added using the automatic synonym expander built into the data mining tool and supplemented with synonyms and abbreviations commonly used by the query developing authors (Supplement 1a).

For precluding automatic retrieval of information entered in the EHR after trial participation, only data registered in EHRs prior to the screening of the trial were used. No site-specific optimizations were added to the query, except for the retrieval of trial participants and periprocedural drug recognition adjustments. To approximate data collection as would have been performed in the trial, the most recent status on any data point before entering the trial was taken. Additionally, drug use data were limited to data registered within a year of enrollment. When no measurement of a variable was found, it was assumed to be absent for the participant.

2.3.2. Manual participant identification, as used in the LoDoCo2 (reference standard)

Trial investigators of the LoDoCo2 trial used two steps to identify trial participants. First, manual screening was performed for eligibility using the EHR files prior to their outpatient clinic visit. Second, patients were interviewed face-to-face to verify eligibility and ask for participation. After providing informed consent, participation in the trial ensued.

2.4. Baseline data extraction methods

2.4.1. Automatic, using text-mining from EHRs

A query was developed to automatically collect data form the EHRs on nineteen variables, which contained information about demography, medical history, procedure history, and drug use as reported in the baseline table of the trials' methods paper (Supplement 1b). For the development of this query the same methods were employed as used for the participant identification query.

2.4.2. Conventional data extraction, as used in the LoDoCo2 trial (reference standard)

In the LoDoCo2 trial, data were collected manually during face-to-face baseline interviews at trial enrollment with participants. Interview data was first recorded as source data on-site and afterward entered into the trial's EDC system.

2.5. Analysis

2.5.1. Participant identification efficiency

For each site, the number of unique patient visits during the trial recruitment period, number of patients automatically identified as potentially eligible, and number of patients enrolled in the trial were recorded and compared to the number of patients enrolled in the actual trial. For both methods, a theoretical yield was calculated based on the patients needed to screen for identification. For determining the yield of the automatic participant identification, the number of enrolled trial participants was used as a proxy as it was not possible to assess how many of the automatically identified potentially eligible patients would have been enrolled retrospectively.

2.5.2. Data collection accuracy

Results of automated EHR text-mining were compared to manually collected trial data on their distributions and accuracy (defined as [true positive data points + true negative data points]/all data points) on an individual patient level. For clarity and to show agreement between EHR vendors, accuracies of the various medical centers were plotted against the overall accuracy in a forest plot.

3. Results

3.1. Participant identification efficiency

A total of 92,466 patients visited the cardiology outpatient clinic of the three study centers during the recruitment period of the LoDoCo2 trial (October 1, 2016 to December 1, 2018). Of these, 568 patients (0.6%) were enrolled in the LoDoCo2 trial (Fig. 3, Table 1).

For the LoDoCo2 trial, all patients visiting the cardiology out-patient clinics were screened on trial eligibility. Automated EHR data screening resulted in a reduction of 73,863 (79.9%) patients that needed to be screened for trial participation. The remaining 18,603 (20.1%) contained 458 of the actual trial participants (82.4% of participants). Further inspection of the 110 (17.6%) trial participants missed by the data mining tool showed that in the automatically retrieved data on one or more inclusion or exclusion criteria were missing (no proof of coronary artery disease [found as a coronary angiography; CT coronary angiography or Coronary Artery Calcium Score]: $n = 38$; no known renal function: $n = 41$; date of previous Coronary Artery Bypass unknown: $n = 41$). Characteristics of missed participants did not differ substantially from identified participants (median difference of all variables 1.6%, IQR 3.1%); values were therefore assumed to be missing at random.

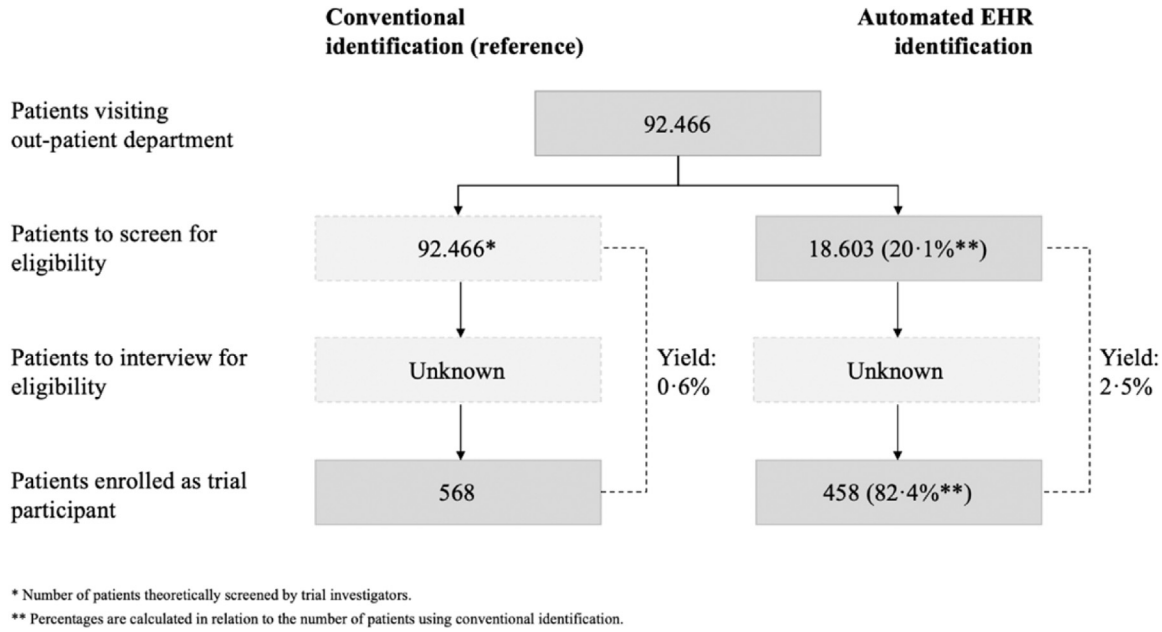


Fig. 3. Eligible patients identified with conventional and automated participant identification.

3.2. Data collection accuracy

Of the 568 trial participants, 540 (95.1%) enrolled trial participants were automatically retrieved on their trial identification number or unique on-site identifier with the data mining tool.

On an aggregate level, availability of baseline characteristics for participants using automated EHR text-mining differed by 2.8% (median; IQR across all variables 0.4–8.5%) with manually collected trial data (Table 2; center-specific distributions are presented in Supplement 2a). Notably larger differences between automated EHR text-mining data and manually collected trial data were found for hypertension (26.2%), antiplatelet therapy (29.1%), and beta-blocker use (24.4%).

On an individual participant level, automated EHR text-mining data showed 88.0% accuracy (median; IQR 84.7–92.8%) when compared to the conventionally collected trial (Table 2; center-specific accuracy is presented in Supplement 2b). Overall, 9.8% of the data extracted from EHRs were false positive (i.e., data on a variable present in EHR data and not present in trial data), and 3.1% false negative (i.e., data on a variable not present in EHR data and present in trial data) (Table 3; for contingency tables of different medical centers see Supplement 2c). Of all data points, positive predictive value was 0.928, negative predictive value was 0.937, sensitivity was 0.806, specificity was 0.827, and F1-score was 0.863 (for test performance scores of individual variables, see Supplement 2d). The lowest accuracies were found for hypertension (62.6%), antiplatelet therapy (68.8%), and beta-blocker use (73.3%). Accuracies for hypertension, antiplatelet therapy, and beta-blocker therapy differed between the participating medical centers, with hypertension ranging from 52.2% to 64.2%, antiplatelet therapy from 60.3% to 86.4% and beta blocker use ranging from 66.4% to 84.7%.

4. Discussion

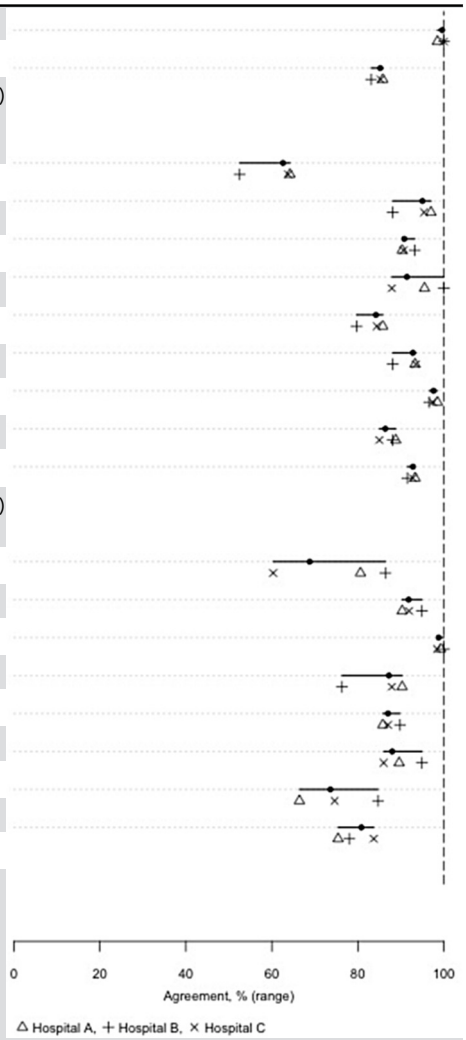
This study shows that it is feasible to use automated EHR text-mining to identify eligible trial participants and collect baseline data. By identifying eligible patients, only

Table 1. Number of patients visiting, eligible, and enrolled per participating medical center

Medical center	Enrollment period	Total no. of patients visiting	No. of trial participants (%)	No. of patients automatically identified as potentially eligible (% of all visiting patients)	No. of trial participants identified as potentially eligible (%; % of participants)
Hospital A	February 1, 2017–October 1, 2018	51,943	169 (0.3)	10,705 (20.6)	151 (1.4; 89.3)
Hospital B	July 1, 2017–October 1, 2018	14,206	69 (0.5)	2,966 (20.9)	65 (2.2; 94.2)
Hospital C	October 1, 2016–December 1, 2018	26,317	330 (1.3)	4,932 (18.7)	252 (5.1; 76.4)
Total		92,466	568 (0.7)	18,603 (20.1)	468 (2.5; 82.4)

Table 2. Distributions and accuracy of baseline variables automatically collected from EHR data compared to trial data

Variable	Trial data.%	EHR data.%	Absolute difference, %	Agreement.%
Sex (Male)	83.6	82.8	−0.8	99.6
Current smoker (Yes)	14.2	13.5	−0.7	85.2
Demographics, median (IOR)			−0.7 (0)	92.4 (7.2)
Hypertension (Yes)	53.6	79.8	26.2	62.6
Diabetes (Yes)	18.4	20.4	2.0	95.0
Insulin-dependent diabetes (Yes)	5.8	13.3	7.5	90.8
Renal function (Not impaired)	91.8	88.5	−3.3	91.4
Prior ACS (Yes)	79.6	84.8	5.2	84.2
Prior PCI (Yes)	87.0	91.9	4.9	92.8
Prior CABG (Yes)	11.5	12.0	0.5	97.6
Atrial fibrillation (Yes)	13.7	8.0	−5.7	86.4
Gout (Yes)	7.3	7.6	0.3	92.8
Medical history, median (IQR)			2 (4.9)	91.4 (6.4)
Antiplatelet therapy (APT) (Yes)	69.2	98.3	29.1	68.8
Oral anticoagulant therapy (OAC) (Yes)	14.8	22.8	8.0	91.8
No APT or OAC (Yes)	0.4	1.1	0.7	98.8
Statin (Yes)	91.8	93.1	1.3	87.2
Ezetimibe (Yes)	23.5	26.3	2.8	87.0
ACE Inhibitor (Yes)	70.3	79.3	9.0	88.0
Beta blocker (Yes)	67.5	91.9	24.4	73.6
Calcium channel blocker (Yes)	27.7	41.5	13.8	80.8
Drug use, median (IQR)			8.5 (14)	87.1 (10)
Overall, median (IQR)			2.8 (8.1)	88 (8.1)



20.1% of the original 92,466 visiting patients had to be screened manually for trial inclusion. In this 20.1%, 82.4% of the participants were present. Data extracted from EHRs showed an average accuracy of 87.1% to the manually collected data of the LoDoCo2 trial.

Several studies have investigated the opportunities of using EHRs for recruitment and data collection in clinical research and trials, but only a few compare EHR data to trial data [17–21]. In general, studies focusing on assessing EHR data quality showed mixed results [10,22–24]. Results from studies focusing on structured EHR data and text-mining in separate EHR components generally showed low yields for EHR quality data [22–24]. A study from 2013 assessed the completeness of structured EHR data to trial eligibility criteria originating from multiple trials, showing that 35% of the patient characteristics derived from the eligibility criteria were available in structured EHR data at the time [23]. In the same year, EHR medication lists were shown to have very broad accuracy (10–90%) [22]. Studies automatically text-mining EHRs integrally, however, reported more favorable results with accuracies comparable to those found in this study [10,24]. In addition, registries based on routinely collected data have been reported to be of high value for trial recruitment and data collection [25].

4.1. Implications for using EHR data in clinical research

When the quality of EHR data extraction is of an acceptable level, it could improve efficacy in trial conduct. As such, EHR data collection would allow the reallocation of resources and a reduction in execution costs [7].

Table 3. Overall contingency table of the accuracy of collected baseline variables

	Trial data, no (%)		
	True[c]	False[d]	Overall
Automatically collected EHR data			
True[a]	3,855 (40.6)	929 (9.8)	4,784 (53.4)
False[b]	299 (3.1)	4,417 (46.5)	4,716 (49.6)
Overall	4,154 (43.7)	5,346 (56.3)	9,500 (100)

[a] Data on a variable present in EHR data
[b] Data on a variable not present in EHR.
[c] Data on a variable present in trial data
[d] Data on a variable not present in trial data

4.1.1. Participant identification efficiency

Using automated EHR text-mining, we were able to identify patients potentially eligible for trial participation. These results are in line with the results found by previous studies [18,19,26]. In participant recruitment, a high positive predictive value using automated EHR participant screening (i.e., most patients screened as positive also enroll in the trial) would maximize efficacy improvements [27]. Our study indicates that automated EHR screening has the potential to identify large numbers of eligible participants in a time-efficient and cost-efficient manner (data not shown).

4.1.2. Data collection accuracy

Since baseline characteristics are not always included in final outcome analysis generally, small errors in these data can be acceptable when counterbalanced by improved efficiency. Incorporation of baseline characteristics measured with error in the analyses would only have an effect on research validity when accuracy is not randomly distributed across intervention groups. If random, it could affect the precision of effect estimates after adjustment [12].

Accuracy of automated EHR data collection depends on the amount of missing data and measurement errors. First, variables collected from data can be missing because they were not recorded or not extracted from the data. Physicians often measure and register only what they consider relevant for delivering care. Consequently, (ordinary) characteristics that are desired in clinical research are not registered [28]. Whether this will lead to problems in identifying patients eligible for trial participation differs per variable and context. Missing data on smoking, for example, will be of less value than missing data on coronary revascularization since clinicians will not always ask about smoking but may be expected to document coronary interventions [29]. These factors make it harder to extract data due to ensuing variability in how characteristics are reported. Substantive knowledge on the topics of data to be extracted is therefore still essential.

Second, EHR data could contain more measurement errors because they were not collected and measured in a standardized format, as is generally done in conventional trial data collection. EHR data can, for example, be hampered in its currency (i.e., stored variables are out of date) due to irregular visits of patients. These remain challenges of the use of EHR data that should be addressed in future research.

Third, relevant information encompassed in the EHR can still be missed due to interindividual differences in reporting or reporting errors (abbreviations, misspelling, synonyms). Improved intelligent text-pattern recognition systems might reduce the risk of missing data.

4.2. Future perspectives

EHR data collection will probably be best used in conjunction with other data collection methods instead of replacing them. In the design of trials, investigators can take automated and manual EHR data collection into account in the design phase of the trial. Our results show that automated EHR screening for eligible patients might result in a somewhat different study population compared to the population currently enrolled. Effects on generalizability should be considered, although the resulting patient population might well reflect a more real-world sample of participants if their characteristics differ from the original study population [18]. Benefits of increased efficiency in the identification of eligible patients might make it easier to enroll patients, and as such, reach the desired number of inclusion faster than with conventional participant recruitment.

5. Limitations of this study

This study combined data from multiple medical centers, all using different EHR software vendors, and shows consistent results for the broad range of systems. Yet, three main limitations should be noted on it.

First, the accuracy of the information on hypertension, antiplatelet therapy, and beta-blockers deviated, notably from collected trial data. Deviation between EHR and trial data was probably due to hypertension being defined as ‘‘using antihypertensive drugs’’ in the LoDoCo2 trial, which was hard to mirror in the EHR search query.

Deviations on drug prescriptions and use variables were mainly attributed to registered timeframes of drugs and insufficient indexing of hospital drug prescription systems by the data extraction tool. Moreover, hospital physicians might not have registered home prescriptions for all patients, adequately deviating results on drugs too.

Second, it was assumed that all patients visiting the outpatient cardiology clinics of the three hospitals were screened conventionally for participation in the LoDoCo2 trial. If this was not the case expected yield of automated participant identification would be overestimated in this study.

Third, our Boolean query was not enhanced with natural language processing algorithms because of the limitations of the employed data mining tool and language-specific limitations. Text-mining was, therefore, interpreted broadly as the ability to automatically extract information from unstructured texts.

6. Conclusions

Data extracted from EHRs using text-mining can be used to identify patients eligible for trial participation and for the collection of baseline characteristics. This method might substantially reduce time and costs related to recruitment and data collection in clinical trials. Whether this premise can be realized depends on whether small accuracy losses are deemed acceptable in the context of the trial that is performed. This study focused on patient eligibility screening and participant baseline data collection; future research is needed to assess the quality of outcome detain EHRs.

Acknowledgments

The authors would like to show their gratitude to Marjan van Doorn (Meander Medical Center) and Erik Badings (Deventer Hospital) for assisting with the data collection for this study.

Appendix A

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.11.014.

References

[1] The European medicines agency working group on clinical trials conducted outside of the EU/EEA. Reflection Paper Ethical GCP Aspects Clin Trials Med Prod Hum Use Conducted Outside EU/EEA Submitted Marketing Au. 2012. Available at www.ema.europa.eu. Accessed January 5, 2020.

[2] U.S. Food and Drug Administration. Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry U. 2018. Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry. Accessed January 5, 2020.

[3] Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated costs of pivotal trials for novel therapeutic agents approved by the US food and drug administration, 2015-2016. JAMA Intern Med 2018; 178(11):1451.

[4] Solomon SD, Pfeffer MA. The future of clinical trials in cardiovascular medicine. Circulation 2016;133:2662−70.

[5] Bentley C, Cressman S, van der Hoek K, Arts K, Dancey J, Peacock S. Conducting clinical trials—costs, impacts, and the value of clinical trials networks: a scoping review. Clin Trials 2019;16: 183−93.

[6] Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? Nat Rev Drug Discov 2017;16(6):381−2.

[7] McClellan M, Brown N, Califf RM, Warner JJ. Call to action: urgent challenges in cardiovascular disease: a presidential advisory from the American heart association. Circulation 20191−11.

[8] Sertkaya A, Birkenbach A, Berlind A, Eyraud J. Examination of Clinical Trial Costs and Barriers for Drug Development. Washington, DC: U.S. Department of Health and Human Services; 2014.

[9] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017;26(1):38−52.

[10] Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016;23(5):1007−15.

[11] Vantongelen K, Rotmensz N, Van Der Schueren E. Quality control of validity of data collected in clinical trials. Eur J Cancer Clin Oncol 1989;25(8):1241−7.

[12] Chan SF, Macaskill P, Irwig L, Walter SD. Adjustment for baseline measurement error in randomized controlled trials induces bias. Control Clin Trials 2004;25:408−16.

[13] Nidorf SM, Fiolet ATL, Eikelboom JW, Schut A, Opstal TSJ, Bax WA, et al. The effect of low-dose colchicine in patients with stable coronary artery disease: the LoDoCo2 trial rationale, design, and baseline characteristics. Am Heart J 2019;218:46−56.

[14] Nidorf SM, Fiolet ATL, Mosterd A, Eikelboom JW, Schut A, Opstal TSJ, et al. Colchicine in patients with chronic coronary disease. N Engl J Med 20201−10.

[15] KPMG. EHR vendor market. 2018. Available at https://nchica.org/wp-content/uploads/2018/10/Eckert-Puls.pdf. Accessed September 26, 2019.

[16] Zorgvisie. Het complete epd-overzicht: welk ziekenhuis heeft welke leverancier? - Zorgvisie. Available at https://www.zorgvisie.nl/epd-overzicht/. Accessed October 1, 2019.

[17] Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013;274(6):547−60.

[18] Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clin Trials 2019; 16:194−203.

[19] Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, et al. Comparison of Electronic Health Record System Functionalities to support the patient recruitment process in clinical trials. Int J Med Inf 2014;83:860−8.

[20] Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;106(1):1−9.

[21] Tissot H, Shah A, Agbakoba R. natural language processing for mimicking clinical trial recruitment in critical care: a semi-automated simulation based on the LeoPARDS trial. Medrxiv 201919005603.

[22] Walsh KE, Marsolo KA, Davis C, Todd T, Martineau B, Arbaugh C, et al. Accuracy of the medication list in the electronic health record—

implications for care, research, and improvement. J Am Med Inform Assoc 2018;25(7):909−12.

[23] Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013;13(1):37.

[24] Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:1−9.

[25] Buccheri S, Sarno G, Fröbert O, Gudnason T, Lagerqvist B, Lindholm D, et al. Assessing the nationwide impact of a registry-based randomized clinical trial on cardiovascular practice. Circ Cardiovasc Interv 2019;12(3):e007381.

[26] Sumi E, Teramukai S, Yamamoto K, Satoh M, Yamanaka K, Yokode M. The correlation between the number of eligible patients in routine clinical practice and the low recruitment level in clinical trials: a retrospective study using electronic medical records. Trials 2013;14(1):426.

[27] Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65:343−349.e2.

[28] Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018;42(11):214.

[29] Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46(5):830−6.