Universiteit Leiden
The Netherlands

An electronic health record text mining tool to collect real-world drug treatment outcomes: a validation study in patients with metastatic renal cell carcinoma

Laar, S.A. van; Gombert-Handoko, K.B.; Guchelaar, H.J.; Zwaveling, J.

Citation

Laar, S. A. van, Gombert-Handoko, K. B., Guchelaar, H. J., & Zwaveling, J. (2020). An electronic health record text mining tool to collect real-world drug treatment outcomes: a validation study in patients with metastatic renal cell carcinoma. *Clinical Pharmacology & Therapeutics*, *108*(3), 644-652. doi:10.1002/cpt.1966

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3181911

Note: To cite this publication please use the final published version (if applicable).

Check for updates

An Electronic Health Record Text Mining Tool to Collect Real-World Drug Treatment Outcomes: A Validation Study in Patients With Metastatic Renal Cell Carcinoma

Sylvia A. van Laar[1], Kim B. Gombert-Handoko[1], Henk-Jan Guchelaar[1] and Juliëtte Zwaveling[1,*]

Real-world evidence can close the inferential gap between marketing authorization studies and clinical practice. However, the current standard for real-world data extraction from electronic health records (EHRs) for treatment evaluation is manual review (MR), which is time-consuming and laborious. Clinical Data Collector (CDC) is a novel natural language processing and text mining software tool for both structured and unstructured EHR data and only shows relevant EHR sections improving efficiency. We investigated CDC as a real-world data (RWD) collection method, through application of CDC queries for patient inclusion and information extraction on a cohort of patients with metastatic renal cell carcinoma (RCC) receiving systemic drug treatment. Baseline patient characteristics, disease characteristics, and treatment outcomes were extracted and these were compared with MR for validation. One hundred patients receiving 175 treatments were included using CDC, which corresponded to 99% with MR. Calculated median overall survival was 21.7 months (95% confidence interval (CI) 18.7–24.8) vs. 21.7 months (95% CI 18.6–24.8) and progression-free survival 8.9 months (95% CI 5.4–12.4) vs. 7.6 months (95% CI 5.7–9.4) for CDC vs. MR, respectively. Highest F1-score was found for cancer-related variables (88.1–100), followed by comorbidities (71.5–90.4) and adverse drug events (53.3–74.5), with most diverse scores on international metastatic RCC database criteria (51.4–100). Mean data collection time was 12 minutes (CDC) vs. 86 minutes (MR). In conclusion, CDC is a promising tool for retrieving RWD from EHRs because the correct patient population can be identified as well as relevant outcome data, such as overall survival and progression-free survival.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Real-world data can provide necessary insights into drug treatment outcomes in clinical practice. The electronic health record (EHR) is one of the potential data sources.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Can efficiency of EHR information extraction on oncologic treatment outcomes be improved by use of the text mining software tool Clinical Data Collector (CDC)?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ Using a test population of patients with metastatic renal cell cancer receiving a range of systemic treatments, we were able to select 99% of the manual population, extract survival data, and structurally stored data as laboratory results with no significant difference to manual review (MR). We were also able to collect comorbidities, cancer-related variables, and side effects with medium to high accuracy. Compared with MR, the use of CDC resulted in a sevenfold reduction of time per patient.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Using CDC, which is a more efficient method for data extraction, continuous treatment evaluation in clinical practice can be facilitated and effectiveness of new drugs in clinical practice can be assessed.

Randomized controlled trials (RCTs) are the gold standard to investigate efficacy of novel drug therapies and, therefore, RCTs are pivotal for drug marketing authorization applications.[1–3] However, in the accelerated approval pathway of the US Food and Drug Administration (FDA) and in the conditional marketing approval pathway of the European Medicines Agency (EMA), new and mostly expensive anticancer drugs are increasingly approved based upon studies with surrogate end points such as progression-free survival (PFS) or objective response rate, and a large part of these studies lack a standard-of-care

[1]Department of Clinical Pharmacy and Toxicology, Leiden University Medical Center, Leiden, The Netherlands. *Correspondence: Juliëtte Zwaveling (J.zwaveling@lumc.nl)

Received February 16, 2020; accepted June 15, 2020. doi:10.1002/cpt.1966

control arm.[4] Consequently, the treatment effect in terms of overall survival (OS) is unclear at approval by the authorities. In addition, novel drugs are usually investigated in a highly selected patient population, which may not be representative for the full cohort of patients who will receive the treatment in clinical practice.[5] This inferential gap between evidence from RCTs and clinical practice can be closed by the use of real-world data (RWD) as complementary information.[1,6–10] These RWD may differ from outcome data from RCTs and may be valuable in assessing the effectiveness of a new drug in daily practice, for example, in patients with specific characteristics, such as older patients or in patients with comorbidities.

An important source for RWD is the electronic health record (EHR).[6,9,11] It contains individual longitudinal patient data collected during routine clinical practice and includes information about patients' demographics, health behavior, vital signs, encounters, laboratory data, medication orders, procedures, imaging, health problem lists, and free-text notes.[12] These free-text notes, in particular, contain very detailed and nuanced information about patients, their illnesses, and treatment trajectory, including efficacy and side effects of drug treatment. However, because these free-text notes are unstructured, they are less suitable for automated information extraction.[13,14] Therefore, manual chart review is still the standard method for data collection from EHRs.[12] Unfortunately, this manual method is laborious, time-consuming, and error-prone,[12,13,15] and thus, not a durable approach for the structural collection of RWD from EHRs. Therefore, more advanced methods are highly warranted.

Natural language processing (NLP) and text mining techniques are advanced methods of information extraction of free-text data.[13,14] Although these methods are promising, they are not yet easily applicable as an alternative method to evaluate the effectiveness of treatments in daily practice. Currently, these techniques are mostly used by a few health care institutions with strong informatics departments, where knowledge of informaticians can be combined with knowledge of clinicians.[16] For example, an NLP pipeline to extract urinary incontinence and erectile dysfunction was developed for patient-centered outcomes of prostate cancer treatment.[17] Additionally, a method combining NLP and machine learning techniques was developed by Sohn *et al*. to collect adverse drug events from psychiatry and psychology medical records.[18] Similar studies were performed for drug-named entity recognition, dosage information, and drug exposure extraction and all these studies were limited to one type of outcome.[16]

The Clinical Data Collector (CDC; CTcue B.V., Amsterdam, The Netherlands) is an NLP and text mining-based tool, which is built to collect structured as well as unstructured data from EHRs and is currently available in hospitals in the Netherlands and Belgium. In contrast to other tools, CDC is designed for usage by medical and pharmaceutical professionals, enabling to easily build queries themselves for information extraction on their topic of interest. Using these queries, only relevant parts of the EHRs are shown and results are directly collected into a dataset, thereby potentially improving the efficiency of retrieval of patient data.[19]

CDC may be a useful extraction tool for retrieving RWD from EHRs. Therefore, we designed a validation study to assess the information extraction of clinical trial parameters from the EHRs by CDC with customized queries. Because we are interested in the effectiveness data of specific oncologic drug treatments, we choose to perform this study in patients with metastatic renal cell carcinoma (mRCC) receiving systemic treatment.

METHODS

In this observational, retrospective validation study, CDC was applied to collect patient characteristics, treatment outcomes, and adverse drug events (ADEs) during drug treatments for mRCC from EHRs. These data were compared with manually obtained data from the EHRs. Patient inclusion, patient characteristics, treatment outcomes, ADEs, and data collection time per patient were evaluated. The study was reviewed by the Medical Ethics Review Committee of the Leiden University Medical Center, who determined that the Medical Research Involving Human Subjects Act (WMO) was not applicable to this study.

Study population

Patients, 18 years and older, with mRCC who received drug treatment with cabozantinib, pazopanib, sunitinib, everolimus, or nivolumab were included in the study. Patients underwent drug treatment between January 2015 and May 2019 in the Leiden University Medical Center, The Netherlands.

Collected variables

Variables that are generally presented in RCTs evaluating new drug therapies in mRCC were collected,[20–22] namely general patient related characteristics (sex, age, length, weight, estimated glomerular filtration rate, alanine aminotransferase (ALAT), and aspartate aminotransferase (ASAT)) and disease-related characteristics (histological RCC subtype and prior nephrectomy) at baseline, including also four common comorbidities (hypertension, cardiovascular comorbidities, diabetes mellitus, and chronic obstructive pulmonary disease) and the International Metastatic Renal cell carcinoma Database Consortium (IMDC) criteria to predict prognostic categories (hypercalcemia, neutrophilia and thrombocytosis, anemia, performance status below 80% Karnofsky, and time from diagnosis to systemic drug treatment below 1 year). Furthermore, treatment outcomes were collected, including tumor progression and OS since the start of treatment and four common ADEs (hand-foot syndrome, liver toxicity, diarrhea, and hypertension).

Manual reference

To create a gold standard, manual chart review was performed by a pharmacist who is experienced in working with the EHR both as healthcare professional and as reviewer. Data were collected from the EHR (HiX, Chipsoft B.V., Amsterdam, The Netherlands), which has no build-in term search and recorded in an electronic case report form (eCRF; Castor EDC, Amsterdam, The Netherlands). For each patient, the time to collect data manually in the eCRF was recorded.

Clinical data collector

CDC is a software tool that is linked with the EHR in the hospital. EHR data are transformed by an application programming interface, to enable structured search using the search engine by medical professionals (users). **Figure 1** shows the communication lines between on-premises isolation platform.

Both the patient population and data points can be defined using CDC queries. Structured data can be extracted from the EHR with specified

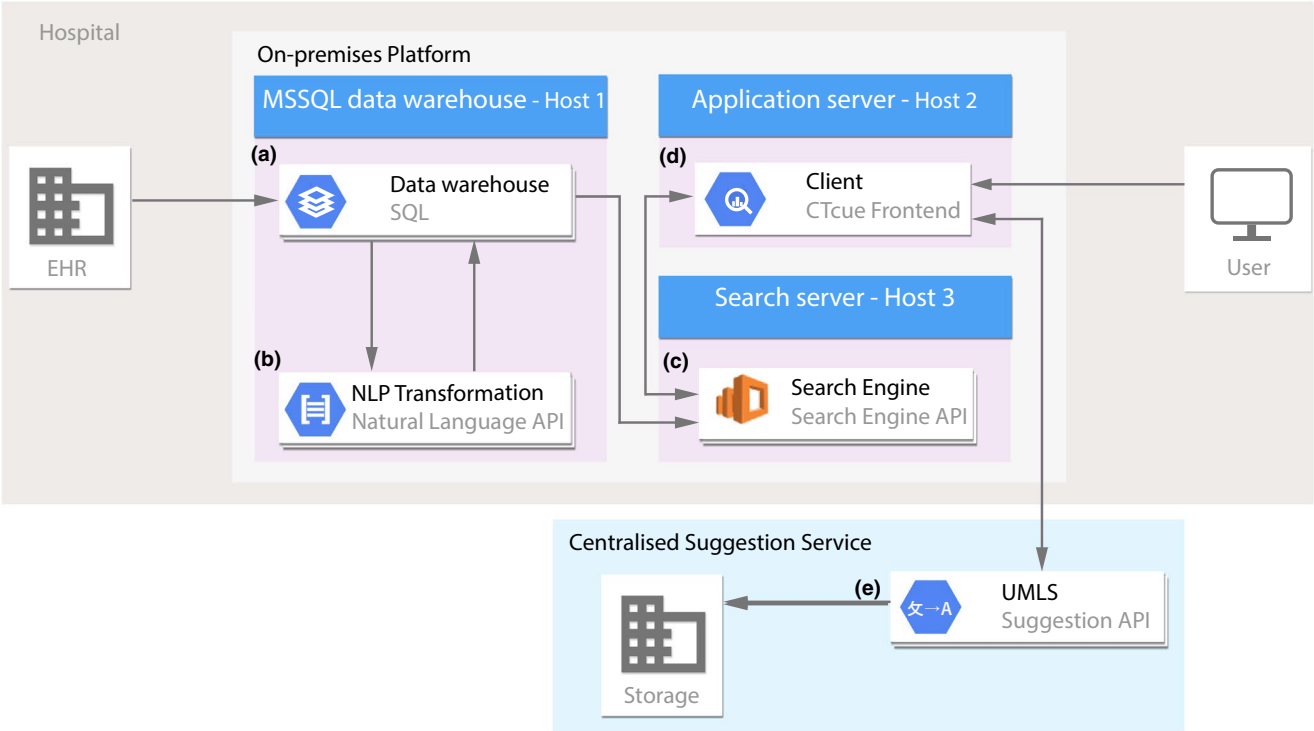

Figure 1 Architecture of the Clinical Data Collector on-premises isolation platform. (**a**) Copy of electronic health record (EHR) data transferred, stored, and cleaned in a local MSSQL Server relational database. (**b**) Natural language processing (NLP) transformation application programming interface (API) pseudonymizes data. (**c**) Search engine is compatible with the structure used in data warehouse. (**d**) Client to build queries by a user. Results window in CDC shows only parts of EHR documents containing defined criteria by user. (**e**) Text mining of (combinations of) keywords is supported by an online thesaurus. [Colour figure can be viewed at wileyonlinelibrary.com]

queries per datatype (e.g., medication requests and laboratory results). Additionally, information extraction from unstructured text is enabled through text mining based on keywords. After running the designed queries, all data are combined in a generic dataset. When a data point is selected, the EHR context is shown, which enables the user to manually validate results. The handling of structured and unstructured EHR data is shown in **Figure 2**. The results can be exported into a CSV-file or XLSX-file.

Queries for patient inclusion and data collection were defined as follows. Patients were included only in CDC for data extraction with both a Diagnosis Treatment Combination (DTC) code for kidney tumors as well as an initial prescription of at least one of the five drug treatments. A DTC is a code used for hospital costs reimbursement in The Netherlands.[23] As both variables were stored as structured data, corresponding structured data queries were applied. The remaining structured data (e.g., age, sex, and laboratory test results) were extracted using these queries as well. For example, the last known measurement result before the start of drug treatment could be automatically selected through linkage with the treatment initiation date. Additionally, queries enabling keyword search were used to select relevant parts of unstructured text in EHRs only. A combination of keywords resulting from the suggestion application programming interface, commonly known synonyms, variants, abbreviations, and typing errors were manually set for this free-text search. In addition, combinations of queries to select structurally stored data and free-text search queries were used to improve recall of some variables. An overview of used queries is provided in **Supplementary File S1**. The completeness of the queries was assessed inspecting the test results section in CDC of 10 random patients and a set of test results was compared with a test set of manual results before finalizing the queries.

After applying patient inclusion criteria using CDC, preselected patients were screened for final inclusion. Subsequently, for data extraction, all variables fully based on structured data were automatically extracted. Variables fully or partially based on unstructured data were manually verified before extraction, using the selected parts of the EHR shown in the results display of CDC, resulting in a semi-automatic extraction procedure. The time spent on final patient inclusion and verification of data was measured for CDC. This was compared with the time that was spent per patient task for manual chart review.

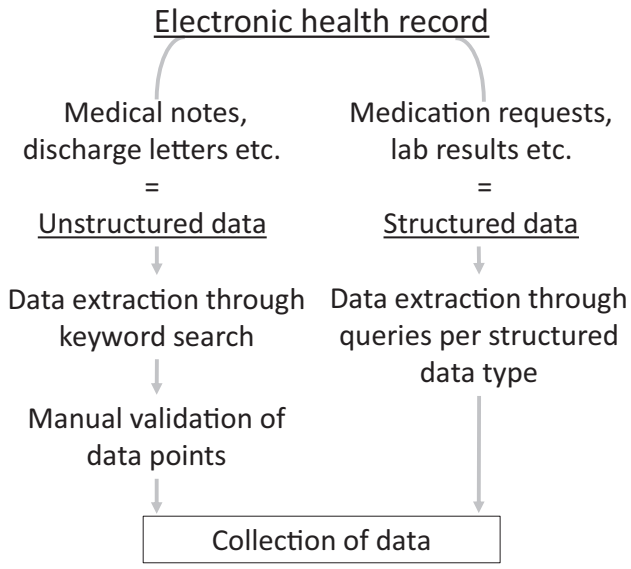

Figure 2 Data extraction approach from structured and unstructured data using Clinical Data Collector.

Analysis and statistics

To establish accuracy of data retrieval, results were compared with manual review (MR). For categorical patient characteristics and ADEs, precision, recall, and F1-scores were calculated. There is no consensus on thresholds for accuracy scores that an information extraction tool should meet. However, we set thresholds for both precision and recall at 90%, to limit the chance on incorrect conclusions when data are used for treatment evaluation. This is in line with thresholds set by Hernandez-Boussard *et al*.[24] Because a part of the IMDC criteria are measurement values, with the answer being a binary question, these will also be analyzed by calculating precision, recall, and F1-score.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{false negatives}}$$

$$F_1 - \text{score} = 2 * \frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{recall}}$$

Next, for all continuous patient characteristics, Bland–Altman plots were composed, to describe agreement between CDC and MR. Per patient, the difference in extracted value was plotted against the mean value of both methods for this patient. In addition, mean differences between data collected using CDC and MR were determined. Kaplan–Meier plots for PFS and OS were composed for all treatments combined. Data were combined because the aim of our study was to validate whether CDC PFS and OS results are equivalent to MR. For PFS, time from start of treatment until significant tumor progression during treatment according to Response Evaluation Criteria in Solid Tumors 1.1[25] was used, or death from any cause. Patients were censored when treatment ended without tumor progression or when patients were still on treatment at the end of inclusion. Furthermore, for OS, time from start of treatment until death from any cause was calculated. Patients were censored when alive at the end of the inclusion period. Because the included patients could have received multiple lines of treatments, patients could occur multiple times in both plots. Statistical analysis was performed in SPSS version 25 (IMB, Armonk, NY).

RESULTS

Patient inclusion

First, we investigated whether CDC was able to trace all patients who met the inclusion criteria. Using inclusion queries in CDC, 133 patients were initially selected based on treatment use and DTC code. Of these, 33 patients were excluded, which resulted in 100 patients included by CDC. For MR, 119 patients were initially selected based on drug prescriptions of cabozantinib, everolimus, nivolumab, pazopanib, and sunitinib in the EHR. These drug treatments represent several treatment lines for mRCC. Of these, 19 patients were excluded, and, therefore, 100 patients were included in the manual dataset. Most of the patients who were excluded in both methods were selected based on a new prescription of follow-up treatment, however, they did not initiate the treatment in the defined inclusion period.

A total of 99 of 100 patients selected using CDC corresponded with the patients included manually. This difference was caused by an incorrectly registered DTC code in the EHR of an unselected patient. **Figure 3** shows the complete patient inclusion flowchart.

Information extraction

Validation parameters were collected and accuracy scores were calculated in order to qualify the usefulness of CDC with respect to manually retrieved outcome data. **Table 1** presents an overview of the collected variables per drug treatment for both methods. First, both MR and CDC identified 175 treatments, of which 174 were identical. The two differences in treatments were due to prescribing errors. One patient did not start treatment with nivolumab according to free-text documentation, which was manually recorded, whereas documented in the structured medication overview and, therefore, extracted by CDC, and vice versa for a treatment of sunitinib. Clear cell RCC was the most frequently reported histological subtype, with 152 patients by MR and 151 patients by CDC. Fourteen patients were manually identified as rarer subtypes, whereas nine remained unclear. CDC reported 6 and 18 patients, respectively. The reported values of other cancer-related variables were also similar. The most reported ADE by MR was liver toxicity (n = 69), however, diarrhea was mostly reported by CDC (n = 51). Hand-foot syndrome was the least reported by both methods (MR: 26; CDC: 19). Furthermore, the number of reported ADEs showed the largest difference between both data retrieval methods for liver toxicity (MR: 69; CDC: 39) and the smallest for hand-foot syndrome (MR: 26; CDC: 19). Further, of all IMDC score parameters used to determine the mRCC prognosis, the incidence of anemia was by far the most reported by both methods (MR: 103; CDC: 105). The reported incidence is quite similar between methods for anemia and thrombocytosis (absolute difference of 0.4% and 0.6%. respectively). Although, an absolute difference of 22% was shown in reported

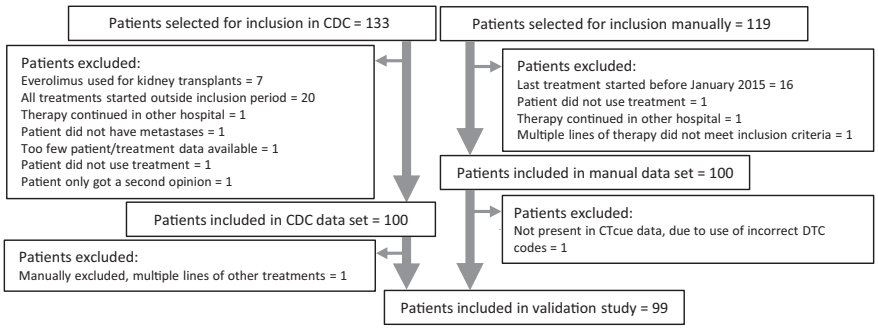

Figure 3 Flowchart of patient inclusion of manual inclusion and inclusion with Clinical Data Collector (CDC). The two approaches yielded patient samples that were very similar and therefore use of CDC is satisfactory for the intended purpose. DTC, Diagnosis Treatment Combination.

Table 1 Collected variables for each treatment per method

	Manual review (n = 175)[a]	Clinical Data Collector (n = 175)[a]
Drug treatment		
Cabozantinib, *n* (%)	27 (15.4)	27 (15.4)
Everolimus, *n* (%)	17 (9.7)	17 (9.7)
Nivolumab, *n* (%)	40 (22.9)	41 (23.4)
Pazopanib, *n* (%)	70 (40.0)	70 (40.0)
Sunitinib, *n* (%)	21 (12.0)	20 (11.4)
Male, *n* (%)	128 (72.7)	129 (73.3)
Cancer-related variables		
Histological subtype of renal cell carcinoma		
Clear cell (%)	152 (86.9)	151 (86.3)
Papillary, *n* (%)	7 (4.0)	3 (1.7)
Sarcomatoid, *n* (%)	3 (1.7)	3 (1.7)
Mixed, *n* (%)	4 (2.3)	0 (0)
Unclear, *n* (%)	9 (5.1)	18 (10.3)
Prior nephrectomy, *n* (%)	114 (65.1)	117 (66.9)
Progression on treatment, *n* (%)	101 (57.7)	98 (56.0)
Death since start treatment, *n* (%)	99 (56.7)	99 (56.7)
Comorbidities		
Hypertension, *n* (%)	91 (52.3)	114 (65.1)
Cardiovascular comorbidities, *n* (%)	43 (24.6)	27 (15.4)
Diabetes mellitus, *n* (%)	39 (22.3)	34 (19.4)
COPD, *n* (%)	12 (6.9, n = 172)	15 (8.6)
Adverse drug events		
Hand-foot syndrome, *n* (%)	26 (14.8)	19 (10.8)
Liver toxicity, *n* (%)	69 (39.2)	39 (22.2)
Diarrhea, *n* (%)	43 (24.4)	51 (29.0)
Hypertension, *n* (%)	64 (36.4)	46 (26.1)
IMDC score parameters		
Hypercalcemia, *n* (%)	28 (16.9, *n* = 166)	24 (14.5, *n* = 166)
Anemia, *n* (%)	103 (59.2, *n* = 174)	105 (60.0, *n* = 175)
Neutrophilia, *n* (%)	32 (20.9, *n* = 153)	40 (24.9, *n* = 162)
Thrombocytosis, *n* (%)	23 (13.4, *n* = 172)	22 (12.8, *n* = 172)
Performance status < 80% Karnofsky, *n* (%)	28 (17.9, *n* = 156)	13 (11.7, *n* = 111)
Time from diagnosis to systemic therapy < 1 year, *n* (%)	89 (50.9)	49 (28.0)

(Continued)

Table 1 (Continued)

	Manual review (n = 175)[a]	Clinical Data Collector (n = 175)[a]
Continuous variables		
Age, years, mean	65.0	65.2
Length, cm, mean	176.2 (*n* = 173)	176.6 (*n* = 169)
Weight, kg, mean	80.6 (*n* = 164)	81.2 (*n* = 148)
ALAT, U/L, median	21 (*n* = 173)	21 (*n* = 174)
ASAT, U/L, median	22 (*n* = 172)	22 (*n* = 174)
eGFR, mL/minute/1.73/m^2, mean	64.9 (*n* = 174)	62.0 (*n* = 155)

ALAT, alanine transaminase; ASAT, aspartate aminotransferase; COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium; U/L, units/Liter.
[a]In case of missing data, number of known variables is presented.

patients who received systemic treatment within a year after diagnosis. A substantial amount of missing data was reported on the IMDC-criteria calcium (MR: 9; CDC: 9), neutrophil (MR: 22; CDC: 13), and performance status (MR: 19; CDC: 64). Finally, the means of all continuous variables were similar. Values for age (years), length (cm), weight (kg), ALAT (U/L), and ASAT (U/L) all differed less than one measurement unit. The reported means for estimated glomerular filtration rate showed a difference of 2.9 mL/minute/1.73 m^2. Moreover, for all variables some missing data was found, however, length (CDC: 6), weight (MR: 11; CDC: 27), and kidney function (CDC: 20) were most prominent.

To assess the quality of data extraction of categorical variables by CDC, the precision, recall, and F1-scores, summarizing both precision and recall, were calculated and presented in **Table 2**. In general, the highest scores on data retrieval were established in cancer-related variables and lowest in ADEs. Besides, results for IMDC-criteria were most diverse with higher scores for continuous structured variables. The highest score for precision of 100% was obtained for sex and platelet levels above normal, and the lowest precision of 39.1% was obtained for performance status. Similar, the highest recall of 100% was reached for sex, platelet levels, and cardiovascular disease, and the lowest score of 63.2% was obtained for hand-foot syndrome.

Outcome parameters were validated by determining PFS and OS. Progression during treatment could be predicted with a precision of 93.1% and recall of 96% by CDC (**Table 2**). In addition, calculated median PFS was 8.90 months (95% confidence interval (CI) 5.38–12.43) vs. 7.59 months (95% CI 5.74–9.44) for CDC vs. MR, respectively (**Figure 4a**), which was not significantly different. Until the seventh month, the curves for PFS overlap, subsequently they split slightly. Death after start treatment was 100% similar extracted by CDC as by MR (**Table 2**) and calculated median OS was 21.72 months (95% CI 18.69–24.75) vs. 21.72 months (95% CI 18.59–24.84), which was equal for both methods (**Figure 4b**). Although CDC reports 77 events with respect to 75 for CDC vs. MR, the curves almost fully overlap.

Bland–Altman plots with mean values of the continuous variables plotted against difference per value were composed to assess

Table 2 Performance scores on collection of categorical variables

	Precision (%)	Recall (%)	F1-score (%)
Sex	100[a]	100[a]	100[a]
Cancer-related variables			
Death since start treatment	100[a]	100[a]	100[a]
Prior nephrectomy	96.5[a]	94.0[a]	95.2[a]
Progression during treatment	93.1[a]	96.0[a]	94.5[a]
Histological subtype of renal cell carcinoma	89.3	88.5	88.1
Comorbidities			
Diabetes mellitus	84.6	97.1[a]	90.4[a]
COPD	91.6[a]	73.3	81.1
Cardiovascular comorbidities	62.8	100[a]	77.1
Hypertension	80.2	64.6	71.5
Adverse drug events			
Diarrhea	81.4	68.6	74.5
Liver toxicity	49.3	87.2	63.0
Hypertension	51.6	71.7	60.0
Hand-foot syndrome	46.2	63.2	53.3
IMDC-criteria			
Thrombocytosis	100[a]	100[a]	100[a]
Anemia	99.0[a]	98.1[a]	98.6[a]
Hypercalcemia	80.1	91.3[a]	85.7
Neutrophilia	90.0[a]	72.9	80.5
<1 year from diagnosis to systematic treatment	53.9	98.0[a]	69.6
Karnofsky performance status < 80%	39.1	75.0	51.4

COPD, chronic obstructive pulmonary disease; IMDC, International Metastatic Renal Cell Carcinoma Database Consortium.
[a]Meet the set threshold for accuracy of 90%.

the quality of continuous data extraction by CDC (**Figure 5**). Because all CIs include 0, differences of means were not significant. Data for age, ALAT, and ASAT showed the best concurrence between both methods.

Extraction time

The total time spent on patient inclusion and information extraction using CDC was 12 minutes per patient, in contrast to 86 minutes spent per patient during MR. This indicates that use of CDC could result in a sevenfold time reduction for the information extraction.

DISCUSSION

This study shows that main treatment outcomes, such as PFS and OS, can be accurately collected using CDC as NLP and text mining software. These most important outcomes met the set standard of 90% for recall and precision. Furthermore, the Kaplan–Meier plots, including time to event, showed no significant differences. Therefore, we conclude that CDC can be adequately applied to retrieve RWD from EHRs in order to add effectiveness data to complement the efficacy data already obtained from RCTs. We conclude that CDC shows to be a technical solution for more consistent and timely data collection.[26] To our knowledge, this is the first study that investigated the use of an information extraction tool to assess drug treatment outcomes in clinical practice.

Of all the extracted categorical patient characteristics, disease as well as drug-related characteristics, prevalence of a prior nephrectomy, thrombocytosis, or anemia before start of treatment met our standard of > 90% for recall and precision. Although not all categorical data met the standard, in general, cancer-related variables and structured IMDC-criteria could be extracted reliably with CDC. Recall and precision were lower for comorbidities, ADEs, and unstructured IMDC-criteria. The differences between both data collection methods may be explained by the characteristics in the EHR for various types of data. First, variables with less variance in free-text registration options in the EHR (e.g., the structured IMDC-criteria), such as laboratory values and cancer-related variables, showed higher accuracy. When data are retrieved using CDC, parts of the EHR are presented containing the predefined keywords. When there is low variety in words used to document variables, chances are higher that all relevant terms are covered in the CDC-queries. In addition, variables that are stable (e.g., histological subtype of a tumor), seem to be more accurately extracted by CDC than variables of a temporary nature (e.g., comorbidities and ADEs). Wang *et al.*[16] already stated that ADE identification is complex, although identification tools such as CDC can be complementary to MR. Additionally, variables registered in the EHR with typing errors could be missed, unless they are specifically entered as a search key. Because real-world oncologic treatment studies, in general, focus on primary outcomes as PFS and OS to study treatment effectiveness, we can accept a larger uncertainty for patient characteristics and adverse events. However, for follow-up studies focusing on these secondary outcome parameters, improvement of queries with the already available software, or advancing CDC by automatizing synonym handling will be beneficial.

The use of CDC resulted in a sevenfold reduction in time for information extraction per patient, therefore, the use of CDC can highly improve the efficiency in retrieving RWD. In the 12 minutes spent per patient, verification of preselected patients and verification of variables was performed. Time spent for preparation of both methods was not taken into account. MR was prepared by constructing an eCRF and for applying CDC, queries were built. Whereof the latter was perceived as more time-consuming, especially the construction of queries for unstructured data. However, these CDC queries can be used repeatedly, for example, in the same patient population at a later moment in time or in other hospitals.

We observed that inconsistencies in the EHR caused differences between both datasets. First, we observed that the information regarding an event in structured data was occasionally not consistent with the description in free-text notes. This led to differences in data retrieval, because some structurally stored variables were extracted by CDC from their dedicated location only, whereas the

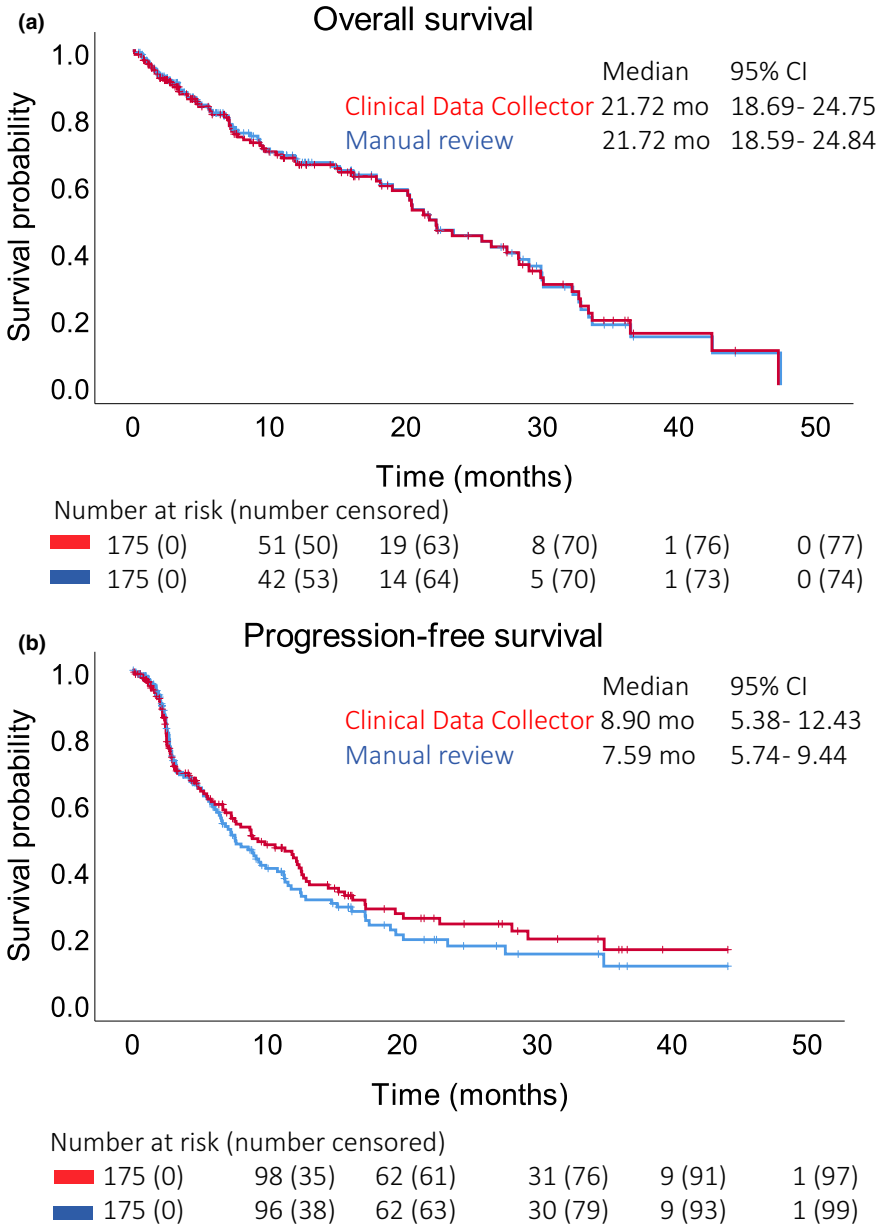

Figure 4 Kaplan–Meier survival plots determined from manual review and Clinical Data Collector data for cabozantinib, everolimus, nivolumab, pazopanib and sunitinib combined. (**a**) Overall survival, (**b**) Progression-free survival. CI, confidence interval.

same variables extracted by MR could be verified consulting free-text notes. For example, errors in using DTC codes could be manually corrected, which was the case in one of our patients. This also applies for inconsistencies between the structured medication list and free-text notes. Because information extraction by CDC was directly linked to the treatment period as deduced from the medication list (structured data), incomplete registration of medication use may influence the extracted values. To illustrate, in our study, start data for drug treatment was not consistent between structured and free-text notes for 72 treatments and differed from 1 day (34 cases) to > 1 year (1 case). Bowman (2013) also described the discrepancies between structured data fields and free-text, for example, for drug dosing instructions.[27] Furthermore, variables such as length, weight, and performance status had to be extracted from unstructured text because these data are not stored in an easily accessible EHR file for CDC. As the CDC extracts information exactly meeting the search criteria only, data can be missed, introducing differences. This may explain the large fraction of missing data and low accuracy scores for the Karnofsky performance status, because these scores are often not literally stated in the EHR. This was also recognized by Hanauer *et al.*,[28] who underlined the variety and errors of numerical values registered in clinical notes. For a few patients, differences in length and weight between CDC and manually retrieved data were remarkably large. We expect these data were subject to measurement errors, typing errors, or were just rough estimations by a physician.

It should be realized that EHR data are RWD, and not a clean data file, such as an eCRF created for research.[1] Therefore, discrepancies

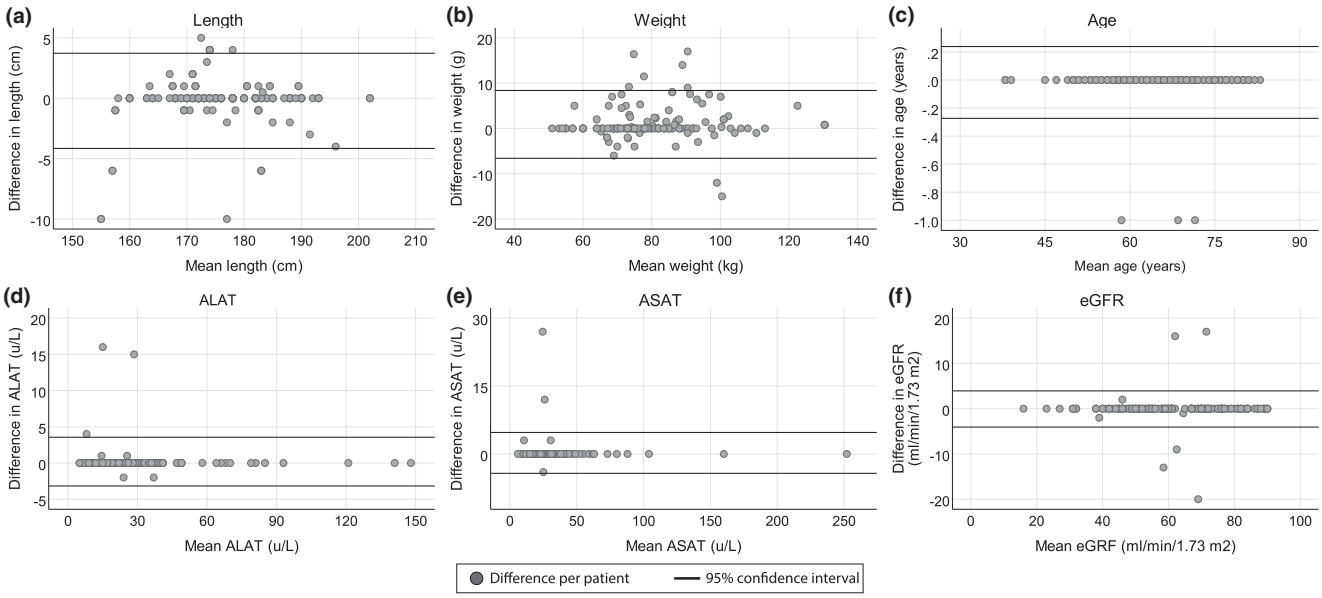

Figure 5 Bland–Altman plots of continuous variables collected using CDC vs. manual with mean difference and 95% confidence interval. (**a**) Length: −0.21 cm (−4.2 to 4.8), (**b**) Weight: 1.1 kg (−6.6 to 8.7), (**c**) Age: −0.17 years (−0.27 to 0.24), (**d**) Estimated glomerular filtration rate (eGFR) 0.22 ml/min/1.73m2 (−5.3 to 5.8), (**e**) Alanine transaminase (ALAT) 0.19 U/L (−3.2 to 3.6), (**f**) Aspartate aminotransferase (ASAT) 0.24 (−4.0 to 4.5).

as well as errors are not completely unavoidable, especially when data are collected in retrospect. Awareness of these errors is necessary when effectiveness of a drug treatment in real life is assessed using data extracted automatically with a tool, such as CDC. However, the results of our study show that despite discrepancies in a few cases, overall, continuous variables were the same between both methods.

This study validated the use of CDC for patient inclusion and data extraction directly on a real-world EHR, on a wide range of variables, as reported in RCTs. Comparison to the gold standard manual reviewed data showed accurate results. A limitation of the study design is that it focused on one type of cancer and its treatments in one Dutch hospital. In addition, in this first study on the accuracy of CDC, data collection was initially performed by one person. Patients were only included for a maximum of ~ 4 years, therefore, not all end points were reached by the time of inclusion ended.

CONCLUSION

We conclude that by using CDC the efficiency of RWD collection can be improved considerably, because patients could be adequately included and treatment outcomes and all structured data could be collected with no significant difference from MR. Although information extraction of unstructured data showed varying results on accuracy, we assume that with some effort suboptimal queries can be optimized for data collection. In the future, these queries can be applied to obtain RWD for several other oncologic drug treatments as well as exported to other centers, which, in particular, can improve efficiency regarding larger and multicenter patient cohorts.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors thank J. de Boer and R. Lakmaker from CTcue B.V. for the provision of Clinical Data Collector for this study and design of **Figure 1**.

FUNDING

This research received no specific grant from any funding agency in public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

H.J.G., J.Z., K.B.G., and S.A.L. wrote the manuscript. J.Z., K.B.G., and S.A.L. designed the research. S.A.L. performed the research. S.A.L. analyzed the data.

© 2020 The Authors. *Clinical Pharmacology & Therapeutics* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

1. Franklin, J.M. & Schneeweiss, S. When and how can real world data analyses substitute for randomized controlled trials? *Clin. Pharmacol. Ther.* **102**, 924–933 (2017).
2. Bothwell, L.E. & Podolsky, S.H. The emergence of the randomized controlled trial. *N. Engl. J. Med.* **375**, 501–504 (2016).
3. Verweij, J. *et al*. Innovation in oncology clinical trial design. *Cancer Treat. Rev.* **74**, 15–20 (2019).
4. Chen, E.Y., Raghunathan, V. & Prasad, V. An overview of cancer drugs approved by the US Food and Drug Administration based on the surrogate end point of response rate. *JAMA Intern. Med.* **179**, 915–921 (2019).
5. Lakdawalla, D.N. *et al*. Predicting real-world effectiveness of cancer therapies using overall survival and progression-free

survival from clinical trials: empirical evidence for the ASCO value framework. *Value Health* **20**, 866–875 (2017).
6. de Lusignan, S., Crawford, L. & Munro, N. Creating and using real-world evidence to answer questions about clinical effectiveness. *J. Innov. Health Inform.* **22**, 368–373 (2015).
7. Skovlund, E., Leufkens, H.G.M. & Smyth, J.F. The use of real-world data in cancer drug development. *Eur. J. Cancer* **101**, 69–76 (2018).
8. Stewart, W.F., Shah, N.R., Selna, M.J., Paulus, R.A. & Walker, J.M. Bridging the inferential gap: the electronic health record and clinical evidence. *Health Aff. (Millwood)* **26**, w181–w191 (2007).
9. Khozin, S., Blumenthal, G.M. & Pazdur, R. Real-world data for clinical evidence generation in oncology. *J. Natl. Cancer Inst*. https://doi.org/10.1093/jnci/djx187.
10. Liu, Q., Ramamoorthy, A. & Huang, S.M. Real-world data and clinical pharmacology: a regulatory science perspective. *Clin. Pharmacol. Ther.* **106**, 67–71 (2019).
11. Sherman, R.E. *et al*. Real-world evidence - what is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
12. Casey, J.A., Schwartz, B.S., Stewart, W.F. & Adler, N.E. Using electronic health records for population health research: a review of methods and applications. *Annu. Rev. Public Health* **37**, 61–81 (2016).
13. Assale, M., Dui, L.G., Cina, A., Seveso, A. & Cabitza, F. The revival of the notes field: leveraging the unstructured content in electronic health records. *Front. Med. (Lausanne)* **6**, 66 (2019).
14. Ford, E., Carroll, J.A., Smith, H.E., Scott, D. & Cassell, J.A. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *JAMIA* **23**, 1007–1015 (2016).
15. Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H.S. & Friedman, C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin. Pharmacol. Ther.* **92**, 228–234 (2012).
16. Wang, Y. *et al*. Clinical information extraction applications: a literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).
17. Hernandez-Boussard, T. *et al*. Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. *AMIA Annu. Symp. Proc.* **2017**, 876–882 (2017).
18. Sohn, S., Kocher, J.P., Chute, C.G. & Savova, G.K. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J. Am. Med. Inform. Assoc.* **18** (suppl. 1), i144–i149 (2011).
19. CTcue <https://ctcue.com/>. Accessed June 25, 2019.
20. Choueiri, T.K. *et al*. Cabozantinib versus sunitinib as initial targeted therapy for patients with metastatic renal cell carcinoma of poor or intermediate risk: the alliance A031203 CABOSUN trial. *J. Clin. Oncol.* **35**, 591–597 (2017).
21. Motzer, R.J. *et al*. Nivolumab versus everolimus in advanced renal-cell carcinoma. *N. Engl. J. Med.* **373**, 1803–1813 (2015).
22. Sternberg, C.N. *et al*. A randomised, double-blind phase III study of pazopanib in patients with advanced and/or metastatic renal cell carcinoma: final overall survival results and safety update. *Eur. J. Cancer* **49**, 1287–1296 (2013).
23. Janssens, P.M.W. Managing the demand for laboratory testing: Options and opportunities. *Clin. Chim. Acta* **411**, 1596–1602 (2010).
24. Hernandez-Boussard, T., Monda, K.L., Crespo, B.C. & Riskin, D. Real world evidence in cardiovascular medicine: assuring data validity in electronic health record-based studies. *J. Am. Med. Inform. Assoc.* **26**, 1189–1194 (2019).
25. Eisenhauer, E.A. *et al*. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
26. Cave, A., Kurz, X. & Arlett, P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin. Pharmacol. Ther.* **106**, 36–39 (2019).
27. Bowman, S. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect. Health Inf. Manag.* **10**, 1c (2013).
28. Hanauer, D.A., Mei, Q., Vydiswaran, V.G.V., Singh, K., Landis-Lewis, Z. & Weng, C. Complexities, variations, and errors of numbering within clinical notes: the potential impact on information extraction and cohort-identification. *BMC Med. Inform. Decis. Mak.* **19**, 75 (2019).