The impact of data extraction and processing on outcomes of research based on routine healthcare data from general practices: an observational study

Melissa Helena Jantien van Essen, Robin Twickler, Yvette M. Weesie, Ilgin G. Arslan, Feikje Groenhof, Lilian L. Peters, Isabelle Bos, Robert A. Verheij

Submitted to: Journal of Medical Internet Research
on: July 22, 2024

Table of Contents

The impact of data extraction and processing on outcomes of research based on routine healthcare data from general practices: an observational study

Melissa Helena Jantien van Essen[1, 2] MSc; Robin Twickler[3] BSc; Yvette M. Weesie[1] MSc; Ilgin G. Arslan[1] PhD; Feikje Groenhof[3] MSc; Lilian L. Peters[3] PhD; Isabelle Bos[1] PhD; Robert A. Verheij[1, 2] Prof Dr

[1]Nivel, Netherlands Institute for Health Services Research Utrecht NL
[2]Tranzo, School of Social Sciences and Behavioural Research Tilburg University Tilburg NL
[3]Department of General Practice and Elderly Care Medicine UMCG, University Medical Centre Groningen Groningen NL

Corresponding Author:
Isabelle Bos PhD
Nivel, Netherlands Institute for Health Services Research
Otterstraat 118
Utrecht
NL

Abstract

Background: Further use of routinely recorded data in electronic health records (EHR) is increasingly more common, for example in epidemiological research. However, data need to be processed and prepared to allow for this further use. Within this process, different choices can be made, which could have significant consequences for research outcomes.

Objective: The aim of this study was to investigate the influence of data processing steps involved in the secondary use of EHR data on research outcomes.

Methods: This study used EHR data from eight Dutch general practices from 2019. These practices contributed data to two research databases: the Academic General Practitioner Development Network (AHON) registry and the Nivel Primary Care Database (Nivel-PCD). Data were extracted and processed using distinct data processing pipelines. This allowed for the evaluation of the impact of different processing methods by comparing the two datasets in a three-step approach: 1) patient demographics, 2) epidemiology of concordant patients, 3) health service utilization of patients with three diagnoses. We compared a number of indicators of similarity between the two databases, including number of contacts, regular consultations and visits, prescriptions, and episodes. Subsequently, for these three diagnoses (diabetes mellitus (DM), urinary tract infection (UTI), cough) we calculated the prevalence, number of prescriptions and number of regular consultations and visits per 1000 patient years. The outcomes were compared by performing two sample t-tests using 99% confidence intervals.

Results: There was a difference in the number of enrolled patients between the two datasets (AHON registry N= 47,517, Nivel-PCD N=44,247). However, the patient demographics were similar. We found differences between all indicator outcomes of the concordant patients in both databases, i.e., the number of contacts, prescriptions and episodes per patient, except for the number of regular consultations and visits (P=.46). Differences in the indicator outcomes varied between the three diagnosis groups, whereas the number of regular consultations and visits was similar between databases for all diagnoses (DM P=<.55, UTI P=.73, cough P=.73)

Conclusions: The results illustrate the importance of awareness of researchers and other users of routine health data of the different steps in processing these data and making them available for research. Data processors should share their knowledge about these choices and researchers and policymakers should invest in their knowledge of this type of metadata. This transparency is all the more important in light of a European Health Data Space and the ever-increasing secondary use of routinely recorded health data. Future research should focus on the role of transparency and joint decision making, to minimize effects of data processing steps and to gain insight into the individual influence of processing steps on research outcomes. This could stimulate a common approach among data processors and researchers resulting in increased data interoperability.

(JMIR Preprints 22/07/2024:64628)
DOI: https://doi.org/10.2196/preprints.64628

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

The impact of data extraction and processing on outcomes of research based on routine healthcare data from general practices: an observational study

Abstract

Background: Further use of routinely recorded data in electronic health records (EHR) is increasingly more common, for example in epidemiological research. However, data need to be processed and prepared to allow for this further use. Within this process, different choices can be made, which could have significant consequences for research outcomes.
Objective: The aim of this study was to investigate the influence of data processing steps involved in the secondary use of EHR data on research outcomes.
Methods: This study used EHR data from eight Dutch general practices from 2019. These practices contributed data to two research databases: the Academic General Practitioner Development Network (AHON) registry and the Nivel Primary Care Database (Nivel-PCD). Data were extracted and processed using distinct data processing pipelines. This allowed for the evaluation of the impact of different processing methods by comparing the two datasets in a three-step approach: 1) patient demographics, 2) epidemiology of concordant patients, 3) health service utilization of patients with three diagnoses. We compared a number of indicators of similarity between the two databases, including number of contacts, regular consultations and visits, prescriptions, and episodes. Subsequently, for these three diagnoses (diabetes mellitus (DM), urinary tract infection (UTI), cough) we calculated the prevalence, number of prescriptions and number of regular consultations and visits per 1000 patient years. The outcomes were compared by performing two sample t-tests using 99% confidence intervals.
Results:
There was a difference in the number of enrolled patients between the two datasets (AHON registry N= 47,517, Nivel-PCD N=44,247). However, the patient demographics were similar. We found differences between all indicator outcomes of the concordant patients in both databases, i.e., the number of contacts, prescriptions and episodes per patient, except for the number of regular consultations and visits (*P*=.46). Differences in the indicator outcomes varied between the three diagnosis groups, whereas the number of regular consultations and visits was similar between databases for all diagnoses (DM *P*=<.55, UTI *P*=.73, cough *P*=.73).
Conclusions: The results illustrate the importance of awareness of researchers and other users of routine health data of the different steps in processing these data and making them available for research. Data processors should share their knowledge about these choices and researchers and policymakers should invest in their knowledge of this type of metadata. This transparency is all the more important in light of a European Health Data Space and the ever-increasing secondary use of routinely recorded health data. Future research should focus on the role of transparency and joint decision making, to minimize effects of data processing steps and to gain insight into the individual influence of processing steps on research outcomes. This could stimulate a common approach among data processors and researchers resulting in increased data interoperability.

Keywords: routine healthcare data; electronic health records; general practice; data processing; data quality; fitness for purpose

Introduction

Secondary use of routine healthcare data is becoming progressively more common, such as the use of electronic health records (EHR) for research and policy making. [1-4] These EHR data are frequently used to report the incidence, prevalence, and the use of health services [3, 5-8], as well as for applications in practice such as decision support, and monitoring and feedback for healthcare providers to improve the quality of care. [9, 10] In the Netherlands, general practice EHR data plays an important role in research and policy making. For example, during the COVID-19 pandemic, EHR data were used to monitor the spread of the disease and the use of health services. [11] Additionally, this valuable information can be used for quality improvement goals without imposing any additional administrative burden on healthcare professionals. [12, 13] Furthermore, from a European perspective, there have been significant developments such as the establishment of a European Health Data Space (EHDS) during the last years. [14, 15] Overall, much good is expected from the developments taking place regarding routine healthcare data.

At the same time, however, there is an ongoing debate about the fitness for purpose of this type of data for secondary use in research and policymaking. [1] Some studies have found EHR data to be accurate and reliable for secondary use, such as the identification of symptoms and diseases, and to be informative of healthcare consumption rates, suggesting no further verification is needed prior to the secondary use of these data. [16, 17] Additionally, studies agree on the value of EHR data and the broad range of purposes that EHR data could be used for. [1, 13, 16, 17] Other studies showed that caution is required when using EHR for secondary purposes, due to the potential introduction of various types of bias, such as selection bias. [1, 4] Since EHR data are primarily aimed at recording the individual patient care as part of the healthcare process, the fitness for purpose of this data for secondary use requires careful consideration. Recent research has emphasized the importance of the concept of fitness for purpose and fitness for use, respectively: data serving intended decision making functions and the ability to get the right information, into the right hands at the right time. [10, 18]

Multiple factors could influence the fitness for purpose of EHR data, such as variation in recording habits of healthcare professionals caused by a high administrative workload [1, 19] and a lack of unity in guidelines, including guidelines for language use in EHR, as well as variations in EHR-software. [4, 5, 20] As several EHR-software systems are commonly used by general practitioners in the Netherlands, the variations of these could influence for example morbidity estimates. [21] In addition, different data extraction and processing methods are likely to affect the final research dataset, for example causing differences within and between general practices. [1]

Verheij et al. visualized the different zones (i.e., care zone, database zone, research zone) present in the data flow for routine healthcare data, see Figure 1. [1] This figure includes the underlying actions (i.e., recording in EHR, extracting data, preparing data for research) and the various actors responsible (i.e. physician, database manager, researcher). It visualizes the various steps taken to process and adapt the data to fit the needs of the researcher, potentially causing bias to unknown levels. In the past, researchers have tried to optimize EHR data for secondary purposes and to limit potential bias, such as confounding bias and selection bias. [22] However, in most studies the focus has been on the completeness of the data, in the research zone. [5, 19, 23]

While previous studies focused on defining the possible sources of bias that can lead to unfit use of EHR data within the 'care zone' and the 'research zone' [1, 4], information regarding the potential effects of the steps taken in the 'database zone' on data fitness for secondary use is lacking. As there can be a multitude of data processors to choose from, a potential beneficiary of this data -e.g., researcher, policy maker, health care professional-, needs to be aware of the possible sources of bias in their respective processing routines. This, however, is not often the case.

Therefore, the aim of this study is to investigate the influence of data processing steps involved in the secondary use of general practice EHR data on research outcomes, such as prevalence rates, between two different EHR-research databases encompassing data from the same eight general practices. This will be investigated by using a selection of indicators to measure the difference between the datasets emanating from the two databases. These indicators are representative for epidemiological and health services utilization studies. Due to the many differentiating steps within these distinct data processing pipelines, we expect these indicator outcomes to be different. These outcomes will demonstrate the extent of the differences, potentially depending on the extent of the data processing. The outcomes of this study will provide valuable insight into the extent of differentiation between EHR databases and possibly contributes to the awareness of users of these sources for secondary purposes.

Methods

Design

This observational study was conducted in the context of the FAIR work packages of three larger COVID-19 related collaborations: COVID-GP [2, 3, 24], Long COVID Mixed Methods [25] and GRIP-3. [26, 27] For these projects pseudonymized EHR data from general practices were stored and combined on the data platform of Statistics Netherlands. For the current study data from eight general practices was used and covered the period from January 1 to December 31 2019. The use of these data allowed for the unique opportunity to evaluate the impact of certain choices made during the different data extraction and processing methods (i.e. database zone steps) by comparing the two datasets in a three-step approach: 1) patient demographics, 2) epidemiology of concordant patients, and 3) health service utilization in three diagnosis groups.

Databases

The datasets used for this study originate from EHR data provided by general practices for two research databases: the Academic General Practitioner Development Network (AHON) registry [28] and the Nivel Primary Care Database (Nivel-PCD). [29] The aim of these registries is to provide insight into epidemiology and health care provided in general practices in the Netherlands. The datasets contain pseudonymized EHR data from 56 general practices participating in the AHON registry located in the North of the Netherlands (approved under number 2020/309) and 363 general practices participating in the Nivel-PCD, located in all regions in the Netherlands (approved under number NZR-00320.087). Eight of these 56 general practices contributed to both databases. The structured data from these shared practices were used to compare research outcomes for the distinctly processed research datasets of AHON registry and Nivel-PCD.

Data extraction and processing pipelines

The AHON registry receives EHR data from a third party that extracts and processes the data. Nivel-PCD receives extracted EHR from the EHR system provider of the general practitioner (GP), after pseudonymization by a third party. This is done according to extraction specifications formulated by AHON registry and Nivel-PCD respectively, one of the distinct steps in the data processing pipelines for the two research databases. The translation into the registry and preparation of the dataset for the researcher contain different steps and choices as well. Table 1 contains an explanation of the differences in processing of each variable between the AHON registry and Nivel-PCD, and the zone in which processing takes place.

Population

The population consisted of all individuals enrolled as patients for at least one quarter in one of the eight shared general practices, of which a subgroup of concordant patients was used for part of the analyses. The concordant patient group consisted of patients present in both databases (83.9% of the total sample), based on their identical identification number as assigned by Statistics Netherlands. This identification number was based on a pseudonym of the social security number for the patients in the Nivel-PCD and for the patients in the AHON registry the identification number was based on a pseudonym of a combination of 3 digits of postal code (PC3), year of birth and sex. The concordant patient group was the group of interest for the analyses of the second step of our three-step approach, the epidemiology of concordant patients, to accurately assess the similarity of data present between patients in the two research datasets, as non-concordant patients will automatically skew the results. The total group of patients – all patients present in the databases – were of interest for the first and third step, namely the patient demographics and the analyses on the health service utilization in three diagnosis groups. This was done to minimize selection bias, as research on health service utilization conducted with EHR data usually does not allow for the filtering of these patients. See Figure 2 for a complete flowchart of the population inclusion.

Indicators of similarity

The datasets include data on demographics of patients, including age, sex, postal code, and registration quarter as well as information on number, type and reason of contacts including consultations and visits and other interventions, diagnoses or symptoms, and number of and indications for prescriptions on patient-level. Information on diagnoses included the International Classification of Primary Care-1 (ICPC-1)-codes and information on prescriptions included the Anatomical Therapeutic Chemical (ATC)-codes. [30, 31] For consultations, visits, and other interventions the insurance claims codes were included, used to record and invoice all activities of the GP, as well as dates. This data is provided in different modules: patient table, contacts table, interventions table and prescriptions table. Each variable within these tables represents a record by the GP. The indicators were operationalized in each of the two databases in separate ways, led by requirements set by the principal investigators of the main project as a preparation for the researcher (Figure 1). The definition and operationalization of these variables or records, as well as in which ‘zone’ the records were processed, are explained in Table 1. When records are processed in different ways, an explanation of the processing step is included for the relevant database, i.e., AHON registry or Nivel-PCD. Variables from the

'research zone' (Table 1), e.g., patient years, prevalence rate and regular consultations and visits, were operationalized in identical ways, as explained in Table 1. Figure 3 presents a schematic overview of the connection between the variables and the zones in which processing takes place. The variables are placed in blue fields representing the actor that produces or processes the variable. The arrows indicate the relationship between the variables. The further down a variable is placed in the overview, the more that variable has been processed.

Table 1. Variables used for analyses (for AHON registry and Nivel-PCD): definitions and processing zones

	General description	AHON registry	Nivel-PCD
Care zone			
ICPC code	ICPC codes are the diagnosis codes based on the International Classification of Primary Care coding system. [30] These codes have been recorded in the EHR by the GP, and are linked to prescriptions, contacts or actions performed by the GP. All general practices in the Netherlands use ICPC-1.	AHON registry contains ICPC codes as recorded by the GP.	Nivel-PCD contains ICPC codes as recorded by the GP.
ATC-code	ATC-codes are the prescription codes based on the Anatomical Therapeutic Chemical classification system for the recording of medication. [31] ATC-codes are recorded in the EHR, by either the GP or via feedback from a different healthcare provider (e.g., the pharmacist).	AHON registry contains ATC-codes as recorded by the GP and does not receive feedback from different healthcare providers.	Nivel-PCD contains all recorded ATC-codes within the GP system. These records can originate either from recording from GPs or from feedback provided by pharmacies. Distinguishing between the two sources is not feasible.
Episodes	Episodes are defined as an episode of illness to which a unique ICPC code is linked. Depending on the EHR system, the end of an episode is automatically recorded in the EHR, and	In the AHON registry, episodes are based on the recorded episodes of care as recorded in the EHR.	Instead of the episodes as defined in the care zone, Nivel-PCD wields the Nivel-PCD 'episode-construct': see 'episode-

	otherwise the GP is required to manually record it. Symptoms or comorbidities can be linked to an episode with the corresponding ICPC-code.		construct' under database zone within this table.
Database zone			
Registration quarter	The registration quarter is the yearly quarter a patient was enrolled at the general practice. The enrollment of each quarter is based on capitation fees that are recorded on a quarterly basis for each patient that is enrolled in the practice during that quarter. The datasets in this study contain information on enrolled patients in 2019, and as such the maximum number of registration quarters per patient is 4. Registration quarters are based on regular yearly quarters, i.e., jan-feb-mar, apr-may-jun, jul-aug-sep and oct-nov-dec.	In the AHON registry, registration duration is based on the date of enrollment of a patient. For this dataset, only fully registered quarters are included. E.g., when a patient enrolls halfway through the quarter, the registration will start from the next quarter.	In the Nivel-PCD, registration duration is based on the date of record of capitation fees, and only patients enrolled for a full quarter are included in this dataset. E.g., when a patient registers halfway through the quarter, the registration will start from the next quarter. In case of a registration present in the first and last quarter of a year, the missing quarters are imputed.
Pseudonymized patient identification number	The patient identification number is a unique number assigned to each patient in a dataset stored on the data platform of Statistics Netherlands. This pseudonymized identifier allows patient information to be linked to other datasets available on the data platform.	In the AHON registry, the patient identification number for patients stored on the data platform of Statistics Netherlands are assigned a patient identification number based on 3 number of the postal code (PC3), year of birth, and sex of the patient.	In the Nivel-PCD, the patient identification number is based on the social security number of the patient. This is the case for patients stored on the data platform of Statistics Netherlands as well as the usual datasets.

		When AHON datasets are not stored on the Statistics Netherlands data platform, a unique patient code based on birth date, sex, patient number within the practice and general practice code is recoded into a unique AHON-patient identification code.	
Prescriptions	In this study, prescriptions are defined as prescribed medications, based on the record of a unique ATC-code on a unique date.	The AHON registry contains all medications prescribed by the GP, including repeat prescriptions.	The Nivel-PCD contains all medications recorded in the EHR. These can be medications prescribed by the GP, including repeat prescriptions, or by a different healthcare provider, which have been submitted as feedback to the GP by either a different healthcare provider or by the pharmacist. The different sources of the prescription are not identifiable in the database. Prescriptions are deduplicated in case of a double record within eight days, for example in case of

			a record by the GP and a record by the pharmacist.
Insurance claims codes	Insurance claims codes are based on the Vektis (insurance claims database) classification system designed to record and invoice all activities of the GP [32] and can be further divided into activities during practice consultations, home visits, other contacts and capitation fee records. These activities can be linked to an ICPC-code on the same day by the data processor, and used to classify and invoice actions such as interventions performed in the general practice, thus providing insight into the invoiced activities of a GP. Insurance claims codes are hence recorded in the care zone, and processed in the database zone.	AHON registry includes all insurance claims codes as recorded by the GP, including the code for recording capitation fees of a patient each quarter.	The Nivel-PCD receives all insurance claims codes as recorded by the GP, however, when a dataset has been requested by the researcher, the codes are filtered by the data processor to include a selection of codes relevant to the study (in agreement with the researchers), on a database zone level. For example, a dataset of the Nivel-PCD usually does not contain the insurance claims code for recording capitation fees, and includes limited interventions carried out by GP practice support.
Contacts	Contacts are defined as moments of contact between a GP and a patient. Contacts are based on unique dates on which an insurance claim code was recorded by the GP, i.e., the maximum number of contacts per patient per day is 1. Insurance claims codes are a classification system designed to record all actions of the GP, and can be further	The AHON registry contains all insurance claims codes present in the general practice EHR system, including interventions carried out by GP practice support.	The Nivel-PCD contains a selection of insurance claims codes relevant for the research.

	divided into practice consultations, home visits and other contacts. The recording of ICPC codes and ATC-codes can be linked to contacts by the data processor, based on date.		
Episode-construct	The episode-construct is an adaptation of the recorded episodes of care as recorded in the EHR by the GP. EHR data of the current year and the two prior years are used to construct the episode. A diagnosis is labeled an episode of illness from the date of diagnosis to the last encounter plus half of the duration of the contact-free interval. [33] The Nivel-PCD thus actively enters an 'end-date' for certain episodes based on this construct, independently of the recording of the GP. Within this construct, when a symptom, such as coughing, is recorded under an episode such as asthma, the symptom will be overruled by the episode.	In the AHON registry, episodes are based on the recorded episodes of care as recorded in the EHR by the GP, and lacks an episode-construct. See 'episodes' in the care zone within this table.	In the Nivel-PCD, episodes are based on the episode-construct, not episodes of care as recorded in the EHR by the GP.
Research zone			
Patient year	The duration of the year that a patient was registered at a general practice. Patient years are calculated based on registration quarters (1-4), thus the minimum number of patient years per patient is 0.25, and the maximum number of	Operationalization of patient years was identical for AHON registry and Nivel-PCD, however differences in outcomes may occur due to differences in the	Operationalization of patient years was identical for AHON registry and Nivel-PCD, however differences in outcomes may occur due to differences in the

	patient years per patient is 1. Patient years are calculated by the researcher.	processing of registration quarters and patient pseudonymization.	processing of registration quarters and patient pseudonymization.
Prevalence rate	The prevalence rate is the total number of patients with a disease existing in the population, in this study per 1000 patient years. The corresponding formula is *number of patients with the record of the ICPC diagnosis code / number of patient years of the population * 1000.* The ICPC codes used for this calculation are contact-ICPC codes. The maximum number of disease cases per ICPC code is 1 per patient. Operationalization of prevalence rate was identical for AHON registry and Nivel-PCD, however differences in outcomes may occur due to the differences in processing of ICPC codes and patient years.	Presence of ICPC diagnosis code was based on the presence in the contacts table of the AHON registry dataset. For differences in the processing of patient years, see 'patient year' in the research zone within this table.	Presence of ICPC diagnosis code was based on the presence in the contacts table of the Nivel-PCD dataset. For differences in the processing of patient years, see 'patient year' in the research zone within this table.
Regular consultations and visits	The regular consultations and visits consist of a subselection of contacts as described under 'Database zone'. The subselection is based on insurance claims codes representing regular consultations and visits [32], and is identical for AHON registry and Nivel-PCD, namely: 12001 – regular consultation, > 20 minutes	All subselected insurance claims codes are present in the AHON registry dataset, and no further data processing takes place on these codes.	All subselected insurance claims codes are present in the Nivel-PCD dataset, and no further data processing takes place on these codes.

	12002 – regular visit, < 20 minutes 12003 – regular visit, > 20 minutes 12010 – regular consultation, < 5 minutes 12011 – regular consultation, > 5 minutes < 20 minutes		

Analyses

Demographic analyses, analyses on epidemiology of concordant patients and analyses on the health service utilization of patients in three diagnosis groups were performed on the Statistics Netherlands remote access platform, where both datasets were uploaded and stored.

Step 1: Demographics

Our first step was to perform demographic analyses on all patients present in the databases. The following demographics were calculated: the total number of unique patients based on the identification number and the total number of patient years, the mean number of patients per practice, the number of patients for each age category (0-4, 5-17, 18-64, 65+ years) and sex. The demographic analyses will provide insight into potential differences present between the two study populations and arranges a relevant perspective on the outcomes of 3, the health service utilization in three diagnosis groups. Due to the differences in pseudonymization methods and registration quarters processing methods, we expect a minor difference in the demographics of the AHON study population and the Nivel-PCD study population. For a definition of patient years and all other variables used for the analyses, see Table 1.

Step 2: Epidemiology of concordant patients

To discover the potential differences present on a patient-level we analyzed the similarity of data present between the two datasets for the concordant patient group. For these analyses, the concordant patient group was used as a way to prevent detection bias and ensure differences in outcome measures present were not overestimated due to differences in the study population. We compared the mean and standard deviation (SD) of indicators of similarity present in both datasets on patient-level: the number of contacts, number of regular consultations and visits, number of prescriptions and number of episodes. For these analyses we merged the available indicators based on the mutual identification number of the patient, and compared the mean and SD for each indicator by performing two sample t-tests and calculating 99% confidence intervals (99%-CI), i.e. the confidence level was set to 99% due to the large number of data. We expected a larger difference in the number of contacts and episodes, due to the more extensive data processing that took place for these variables (Table 1), especially due to the presence of the episode construct for Nivel-PCD episodes. For the number of regular consultations and visits we expected no significant differences as the data extraction and processing on these insurance claims codes was largely identical for the two databases. Analyzing the healthcare consumption of patients based on a sub-selection of relevant insurance claims codes, as opposed to a non-specific selection, i.e., all insurance claims codes, is more representative of research conducted with EHR data, as this is a more commonly

used method among researchers.

Step 3: Analyses on the health service utilization in three diagnosis groups

Subsequently, a set of indicators was selected, providing a way to compare the research outcomes of both datasets for different diagnoses. The purpose of these analyses is to observe the possible effect of the different data extraction and processing pipelines that the two datasets have been subject to. Hence the analyses were performed for all patients present in the datasets.

Selection of indicators

The indicators were selected by a research team with expertise in research conducted with routine healthcare data and data processors on the two databases. With this set of indicators, we had the intention to be representative for the research that is generally conducted with these datasets. [3, 28, 29, 34] When selecting the set of indicators, the selection process was based on including the diagnosis of a chronic or long-term condition, an acute condition, and a symptom with a high prevalence within Dutch general practices and with a high disease burden. Additionally, we were cautious that the selection of indicators utilized all available data provided by the databases, to ensure relevant potential differences present in the data were detected in the outcomes. We included patients with a Diabetes mellitus (DM), urinary tract infection (UTI) and cough diagnosis. For each of these diagnosis groups we calculated the total number of patients, by including the patients with a relevant ICPC code recorded in the EHR, namely T90 for DM, U71 for UTI and R05 for cough. The prevalence rate per 1000 patient years was calculated by dividing the number of patients with one of these ICPC diagnosis codes recorded by the number of patient years of the population, times 1000. Additionally, we calculated the number of regular consultations and visits these patients received per 1000 patient years, and the number of prescriptions per 1000 patient years. This was done by selecting a group of ATC-codes for each diagnosis as indicated by the pharmacotherapeutic compass, a Dutch online reference book that provides independent pharmaceutical information for medical professionals: A10A and A10B for DM, G03C, J01C, J01D, J01E, J01G, J01M and J01X for UTI and R05C, R05D, R05X and R06A for cough. [35]

The analyses were performed for each diagnosis group within each dataset separately. For definitions of patient years, prevalence rate and prescriptions, see Table 1. For each diagnosis group we compared the mean indicator outcome by performing two sample t-tests and calculating the 99% CI. All analyses were performed using R [version 4.2.3] and RStudio [version 2022.02.1+461 “Prairie Trillium”].

Ethical considerations

Ethical approval for this study was waived by the medical ethics committee of the University Medical Centre Groningen (reference number: 2020/309). The use of EHR data is permitted under certain conditions by Dutch law both for the data from the general practice registration network (AHON) and Nivel-PCD. According to this legislation, neither obtaining informed consent from patients nor approval by a medical ethics committee is obligatory for these types of observational studies, containing no directly identifiable patient data (art. 24 GDPR Implementation Act jo art. 9.2 sub j GDPR). For Nivel-PCD, the project has been approved by the relevant governance bodies of Nivel-PCD under no. NZR-00320.087.

Results

Population and demographics characteristics

The population included in this study consisted of all patients registered for at least one quarter at the general practice in 2019 at one of eight general practices mutually present in the AHON registry and the Nivel-PCD, resulting in a total of 49,907 patients, of which 47,517 patients were present in the AHON registry database and 44,247 patients were present in the Nivel-PCD. A subgroup of this population was made, resulting in a group of 41,857 patients present in both datasets: the concordant patients. We had no insight into the number of patients in the general practices before extractions from the EHR system took place. Table 2 provides an overview of the demographic characteristics of all patients combined. The patient demographics age category and sex were similar in both datasets. There was a difference in the total number of unique patients (N=47,517 vs. N=44,247), the number of patient years (N=46,400 vs. N=43,100) and the mean number of patients per practice (N=5,940 vs. N=5,531), the latter stemming largely from one outlier practice.

Table 2. Demographic characteristics of the study population (all patients, N = 49,907).

	AHON registry N=47,517	Nivel-PCD N=44,247
	N(%)	N(%)
Patient years	46,400, 0.977 per patient	43,100, 0.974 per patient
Patients per practice, mean	5,940[a]	5,531[a]
Age category		
0 – 4 years	2,239 (4.7)	2,093 (4.7)
5 – 17 years	7,382 (15.5)	6,917 (15.6)
18 – 64 years	27,507 (57.9)	25,684 (58.0)
65+ years	10,389 (21.9)	9,553 (21.6)
Sex		
Male	23,471 (49.2)	21,968 (49.6)
Female	24,046 (50.6)	22,279 (50.4)

[a] Difference in mean number of patients per practice largely due to one outlier.

Similarity in epidemiology of concordant patients

The number of contacts, regular consultations and visits, prescriptions and episodes were analyzed per patient for the concordant patients in each dataset. By comparing these outcomes, statistically significant differences were obtained between the AHON registry and Nivel-PCD. All differences, except for the difference in number of regular consultations and visits (P=0.57), were significant. In the AHON registry the mean number of contacts recorded per patient was 8.58 (SD 10.10), while in the Nivel-PCD this average was 7.40 (SD 9.02). There was no difference between the average number of regular consultations and home visits for the

patients between AHON registry and Nivel-PCD (*P*=.46), with an average number of 4.33 (SD 5.67) and 4.30 (SD 5.65) respectively. The number of prescriptions was on average 6.75 (SD 11.30) prescriptions per patient in the AHON registry and 5.90 (SD 9.45) prescriptions per patient in the Nivel-PCD (*P*<.001). The number of episodes was significantly lower for the patients in the AHON registry: 1.61 (SD 1.73) episodes per patient in 2019, while patients in the Nivel-PCD had a mean number of 3.74 (SD 3.67) episodes in 2019 (*P*<.001). See Table 3 for the outcomes of all indicators of similarity of the concordant patient group.

Table 3. Epidemiology of the concordant study population.

	AHON registry (N=41.857)	Nivel-PCD (N=41.857)	Statistical differences between AHON registry and Nivel-PCD
	Mean (SD)	Mean (SD)	*P*-value, 99%-CI
Number of contacts per patient	8.58 (10.10)	7.40 (9.02)	<.001, 1.36 – 1.02
Number of regular consultations and visits per patient	4.33 (5.67)	4.30 (5.65)	.46, -0.07 – 0.13
Number of prescriptions per patient	6.75 (11.30)	5.90 (9.45)	<.001, 0.03 – 0.001
Number of episodes per patient	1.61 (1.73)	3.74 (3.67)	<.001, 0.09 – 0.10

Analyses on health service utilization in three diagnosis groups

In step 3 we analyzed the prevalence rate, the number of prescriptions and the number of regular consultations and visits per 1000 patient years for three different diagnosis groups: DM, UTI, and cough. There was a statistically significant difference in the prevalence rate of the UTI (*P*<.001) and cough diagnosis groups (*P*<.001) (Table 4). The number of prescriptions per 1000 patient years was significantly different for the DM diagnosis group (*P*<.001) and the UTI diagnosis group between the two databases (*P*=.009). For the cough diagnosis group there was a low number of prescriptions in both databases, and the difference between the two datasets was not statistically significant (*P*=.014). The regular consultations and visits did not significantly differ for the three diagnoses. All P-values and 99% confidence intervals can be found in Table 4.

Table 4. Indicator outcomes for three diagnoses.

	AHON registry (N=47,517) (patient years=46,400	Nivel-PCD (N=44,247) (patient years=43,100)	Statistical differences between AHON registry and Nivel-PCD

	)		
	N	N	*P*-value, 99%-CI
Diabetes mellitus (T90) diagnosis			
Number of patients with ICPC T90 record (patient years, patient years per patient)	2,979 (2,936, 0.986)	2,787 (2,728, 0.979)	N/A[a]
Prevalence rate	64.2	64.7	.78, -0.005 – 0.004
Number of prescriptions per 1000 patient years	447.0	377.4	<.001, 0.47 – 1.68
Number of regular consultations and visits per 1000 patient years	498.1	529.6	.55, -0.93 – 1.49
Urinary tract infection (U71) diagnosis			
Number of patients with ICPC U71 record (patient years, patient years per patient)	2,960 (2,927.25, 0.977)	3,005 (2,961.25, 0.977)	N/A
Prevalence rate	63.8	69.7	<.001, -0.01 – -0.002
Number of prescriptions per 1000 patient years	138.1	135.0	.009, 0.003 – 0.40
Number of regular consultations and visits per 1000 patient years	722.1	823.3	.73, -0.97 – 1.27
Cough (R05) diagnosis			
Number of patients with ICPC R05 record (patient years, patient years per patient)	2,404 (2,356.5, 0.980)	2,718 (2,663.25, 0.980)	N/A
Prevalence rate	51.8	63.1	<.001, 0.007 – 0.02
Number of prescriptions per 1000 patient years	50.8	51.0	.014, -0.008 – 0.34
Number of regular consultations and visits per 1000 patient years	426.4	536.7	.73, -1.05 – 1.38

[a] N/A: not applicable.

Discussion

This study investigated the influence of data extraction and processing of routine healthcare data on research outcomes, by comparing indicators of similarity based on EHR data from eight general practices that were composed using two different data processing methods. Regardless of the identical origin of the data, the different data extraction and processing pipelines, and the choices made by several actors, i.e., the data processor, the researcher, etc., during different stages of these processing methods by each database, resulted in different indicator outcomes. Our results show more substantial differences when data has been more extensively processed, and no significant differences when processing was minimized.

The value of EHR data is increasingly recognized, and the need for reliable and valid data is widely acknowledged as well. However, our findings highlight the need for transparency regarding these steps and the motives behind them, as well as the need for adequate metadata describing these choices. [36, 37] For example, these choices often originate from the inclination to improve on the validity of the outcomes. It also highlights the fact that the issue of interoperability does not stop with uniform EHR systems or standardized coding or ontologies. [36, 38]

The results support the expectation that data extraction and processing choices can affect research outcomes and highlight the relevance of adopting transparency in the approach to obtaining, processing, analyzing but moreover interpreting the data, when aiming for appropriate quality. Additionally, it shows the complexity of data processing and coordinating certain definitions of variables between data processors and researchers. Furthermore, researchers conducting future studies with EHR data should be mindful of data processing choices made, and data processors should share their knowledge about these choices. Additionally, users of these data such as researchers and policymakers should invest in their knowledge of this type of metadata, as transparency about these systems is becoming increasingly important in light of an EHDS.

Principal Results

First, we compared the demographic characteristics of all patients in the study population and noted a difference in the number of unique patients, patient years and mean number of patients per practice. This may be explained by the different pseudonymization methods, as the AHON registry uses the PC3 postal code, year of birth and sex of the patient and the Nivel-PCD uses the pseudonym of the social security number of the patient. After pseudonymization the patient data was uploaded, stored, and combined on the data platform of Statistics Netherlands. Previous research shows similar procedures result in a loss of coverage of linked patients. [39] The difference in patient years may be caused by the data processing of the registration quarters in each database, as each databases' patient years are based on differentiating dates of registration quarters, and the occurrence of imputation of registration quarters takes place for Nivel-PCD only.

Second, we compared the similarity of the epidemiology of the concordant patient group. There was a statistically significant difference between AHON registry and Nivel-PCD in the average number of contacts, prescriptions, and episodes per patient. This implies that differences occur on the level of the datasets as a whole, as seen from the demographic analyses results, as well as for the exact concordant patients. We conclude these differences occur due to different data extraction and processing methods and additionally, due to the impact of choices on the

analyses. For example, specific selection of insurance claims codes for regular consultations and visits show no significant result, implying that thorough coordination of variables may evade the effects of different data processing methods. The difference in number of episodes may be attributed to the Nivel-PCD episode construct. The Nivel-PCD episode construct algorithm takes all episodes as recorded by general practitioners into account, as well as contacts with the GP and prescriptions as recorded in the EHR. Additionally, this construct includes episodes that were started at the end of the previous year and continues these into the next year. [33] For AHON registry diagnoses based on the episode ICPC, these episodes are not included. The episode construct hence possibly increases the number of episodes compared to the number of episodes based on general practice records, which is in line with the results of this study.

Lastly, we investigated the effects of the different data extraction and processing methods on a selection of indicator outcomes, comparable to real world research conducted with these datasets, by analyzing the health service utilization in three diagnosis groups. For these indicators we found a significant difference in a majority of the outcomes, with the exception of regular consultations and visits. Nivel-PCD and AHON registry have different exclusion steps for the insurance claims codes in preparation for a research dataset, resulting in a large difference in the total number of insurance claims codes and the type of codes included. The sub-selection of these claims codes for regular consultations and visits are specifically coordinated, resulting in no significant difference in the number of regular consultations and visits between the two databases. When the same definition is used for regular consultations and visits, and little to no processing steps have taken place on the variables used, no significant differences are seen between the two databases. This highlights the importance of clear specification when analyzing data. [37] Interpretation differences can occur if insurance claims codes in general were analyzed and as data are increasingly being shared, for example between countries, researchers without the knowledge of certain healthcare systems or data processing methods can unintentionally present biased results. Meta-information on data extraction and processing choices as well as research methods could be a solution.

The prevalence rate was significantly different for the UTI diagnosis group and the cough diagnosis group, but not for the DM diagnosis group. This may be attributed to the processing that takes place for ICPC codes within the contacts table. This table includes the ICPC codes of patients who have visited the GP for this specific diagnosis. Patients with chronic illnesses may visit the GP more frequently for their diagnosis compared to patients with an acute illness, which may diminish the effect of processing, as the maximum number of disease cases for the prevalence rate was one per patient. Additionally, differences may be attributed to the differences in processing of patient years and the pseudonymization process.

In this study, the results of the indicators are dependent on the dataset that is used to answer the research questions. The interpretation of these outcomes is relevant because research outcomes are often used for purposes such as policymaking and feedback information to healthcare professionals, and the approach of interpreting research outcomes when handling imperfect data is thus consequential. Third parties using EHR data for secondary uses should therefore not dismiss the value EHR data has to offer, such as the broadness of research outcomes available, as demonstrated in this study, but rather focus on improving the manner in which these data are handled. The criticality of the interpretation of research outcomes may be different for research outcomes based on trends over time. In other words, when the data extraction and processing methods remain identical for several datasets over the years, and

data processors and researchers focus on data robustness, outcome measures on trends over time might remain reliable. This should be explored in future research.

Comparison with Prior Work

The differences found in the outcomes of the indicators in all three steps highlight the necessity of transparency and joint decision making for and with the knowledge of researchers on the dataset that is being used. Instructions on the fitness for purpose and the data quality can be included in the documentation, and clear communication between data processors and researchers is crucial for the interpretation of researchers and policymakers on the results of their study. The outcomes of this study suggest that frameworks to improve fitness for purpose could prove to be a necessary tool in analyzing and interpreting the data. Previous research has resulted in a data quality assessment framework to improve the quality of the datasets that are used for secondary purposes such as research, but does not elaborate on the effects processing steps can have on research outcomes. [40]

Differences in outcomes that occur due to data processing emphasize the need to make joint decisions regarding data processing pipelines, as this may increase interoperability, for example between research databases. To achieve this, documentation regarding this process is essential and the need for detailed meta-information is crucial in this type of research. Interoperability also requires collaboration within and between data processors and researchers. [41] This cooperation will lead to better interpretations of the research conducted with these types of data, and previous research concludes there are benefits to be gained from research on optimal common standards. [42] Similar common approaches have been recommended to improve data quality and has resulted in a harmonized data quality assessment framework. [43] To stimulate interoperability and increase data quality [44], frameworks such as these should become common practice before analyzing EHR data.

Limitations

A limitation of this study is the choice of diagnoses (i.e., DM, UTI and cough) with high prevalence rates, possibly making the indicator outcomes not applicable to smaller diagnostic groups, for example of rare conditions. [45] The choice for these diagnoses was made to make sure to include a larger number of patients per diagnosis as the number of concordant general practices in the databases was small. An additional limitation is the lack of details into the data extraction and processing steps that were taken for each database. This was due to the fact that this study started with the end products, i.e., the research datasets, as opposed to the unprocessed data coming directly from the EHR systems. Despite the limitations, this data was prepared for research irrespective of this study, which decreases bias in the preparation methods of the extraction and data processing for these research datasets. Moreover, this study appears to be the first study to compare data processing methods with concordant general practices and hence contributes to gaining insight into the influence of these methods on research outcomes based on EHR data.

Conclusions

In conclusion, routine healthcare data such as EHR data from general practices offer a broad spectrum of applications and the secondary use of these data are ever increasing. Moreover, the results show the impact of data processing steps and analysis choices on the beforementioned indicator outcomes and the necessity of transparency between the knowledge of data processors regarding these choices, and the knowledge of researchers of this type of metadata. Researchers and policymakers should be cautious with the secondary use of EHR data,

especially with regards to the interpretation of research outcomes. Future research should focus on this transparency and the benefits of using a data quality framework intended to minimize effects of data processing steps, and on gaining more insight into the individual influence of different processing steps on different research outcomes. This could stimulate a common approach among data processors and researchers and thus increase interoperability, which is all the more important with regards to developments such as EHDS and the ever-increasing secondary use of routinely recorded health data.

Acknowledgements

This research has been conducted using the AHON registry and Nivel-PCD. The AHON registry and Nivel-PCD are ongoing longitudinal databases which aim to give insight into medical care provided in general practices in the North of the Netherlands and the Netherlands, respectively. Starting in 1998 and 1970 respectively, pseudonymized patient records have been added to the AHON registry and Nivel-PCD multiple times a year from a growing number of participating general practices. Patients are informed by their general practices via folders, posters, and websites of their participation in the AHON registry and Nivel-PCD. Individual patients can fill out an opt-out form in order to not have their data recorded in the AHON registry or Nivel-PCD.

The Netherlands Organization for Health Research and Development (ZonMW) funded this study: “Changes in the Use and Organization of Care in General Practices and Out-of-hours Services: Lessons Learned from the COVID-19 Pandemic” (10430022010006) and the “General Practice Research Infrastructure Pandemic Preparedness Program” (GRIP3) (10430112110001). The funder played no role in the study design, data collection, data analysis and interpretation, or writing of this manuscript.

Conflicts of Interest

None declared.

Abbreviations

AHON: Academic General Practitioner Development Network (Academische Huisartsen Ontwikkel Netwerk)
ATC-codes: Anatomical Therapeutic Chemical codes
DM: Diabetes mellitus
EHDS: European Health Data Space
EHR: electronic health records
GP: general practitioner
ICPC-1 codes: International Classification of Primary Care-1 codes
Nivel-PCD: Nivel Primary Care Database
UTI: urinary tract infection

References

1. Verheij, RA, et al., Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res; 2018. 20(5): p. e185. PMID:29844010
2. Ramerman, L, et al., The use of out-of-hours primary care during the first year of the COVID-19 pandemic. BMC Health Services Research; 2022. 22(1): p. 1-9. PMID:35597939
3. Rijpkema, C, et al., Care by general practitioners for patients with asthma or COPD during the COVID-19 pandemic. npj Primary Care Respiratory Medicine; 2023. 33(1): p. 15. PMID:37031214
4. Mc Grath-Lone, L, et al., What makes administrative data research-ready? : A systematic review and thematic analysis of published literature. International Journal of Population Data Science; 2022. 7(1). PMID:35520099
5. Arslan, IG, et al., Incidence and prevalence of knee osteoarthritis using codified and narrative data from electronic health records: a population-based study. Arthritis Care & Research; 2022. 74(6): p. 937-944. PMID:35040591
6. Violán, C, et al., Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. BMC public health; 2013. 13: p. 1-10. PMID:23517342
7. Institute for Health Metrics and Evaluation, How we collect data. IHME health data website; 2024. Data tools and practices.
8. World Health Organization, Health service data. WHO website; 2024. Data collection tools.
9. Mbizvo, GK, et al., The accuracy of using administrative healthcare data to identify epilepsy cases: a systematic review of validation studies. Epilepsia; 2020. 61(7): p. 1319-1335. PMID:32474909
10. Barbazza, E, et al., Optimising the secondary use of primary care prescribing data to improve quality of care: a qualitative analysis. BMJ open; 2022. 12(7): p. e062349. PMID:35863830
11. Madhavan, S, et al., Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States: a perspective from 15 academic medical centers. Journal of the American Medical Informatics Association; 2020. 28(2): p. 393-401. PMID:33260207
12. Horrocks, S, et al., Accuracy of routinely-collected healthcare data for identifying motor neurone disease cases: a systematic review. Plos one; 2017. 12(2): p. e0172639.
13. Dash, S, et al., Big data in healthcare: management, analysis and future prospects. Journal of Big Data; 2019. 6(1): p. 54. PMID:28245254
14. The European Health Data Space (EHDS). EHDS website; 2024. What is the European Health Data Space (EHDS)?
15. Marcus, J. Scott, et al. The european health data space. IPOL | policy department for economic, scientific and quality of life policies, European Parliament Policy Department studies; 2022. DOI:10.2139/ssrn.4300393
16. Harper, C, et al., Comparison of the accuracy and completeness of records of serious vascular events in routinely collected data vs clinical trial–adjudicated direct follow-up data in the UK: secondary analysis of the ASCEND randomized clinical trial. JAMA network open; 2021. 4(12): p. e2139748-e2139748. PMID:34962561
17. Ta, CN, et al., Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. Scientific data; 2018. 5(1): p. 1-17. PMID:30480666
18. Barbazza, E, Klazinga, NS, and Kringos, DS, Exploring the actionability of healthcare performance indicators for quality of care: a qualitative analysis of the literature, expert

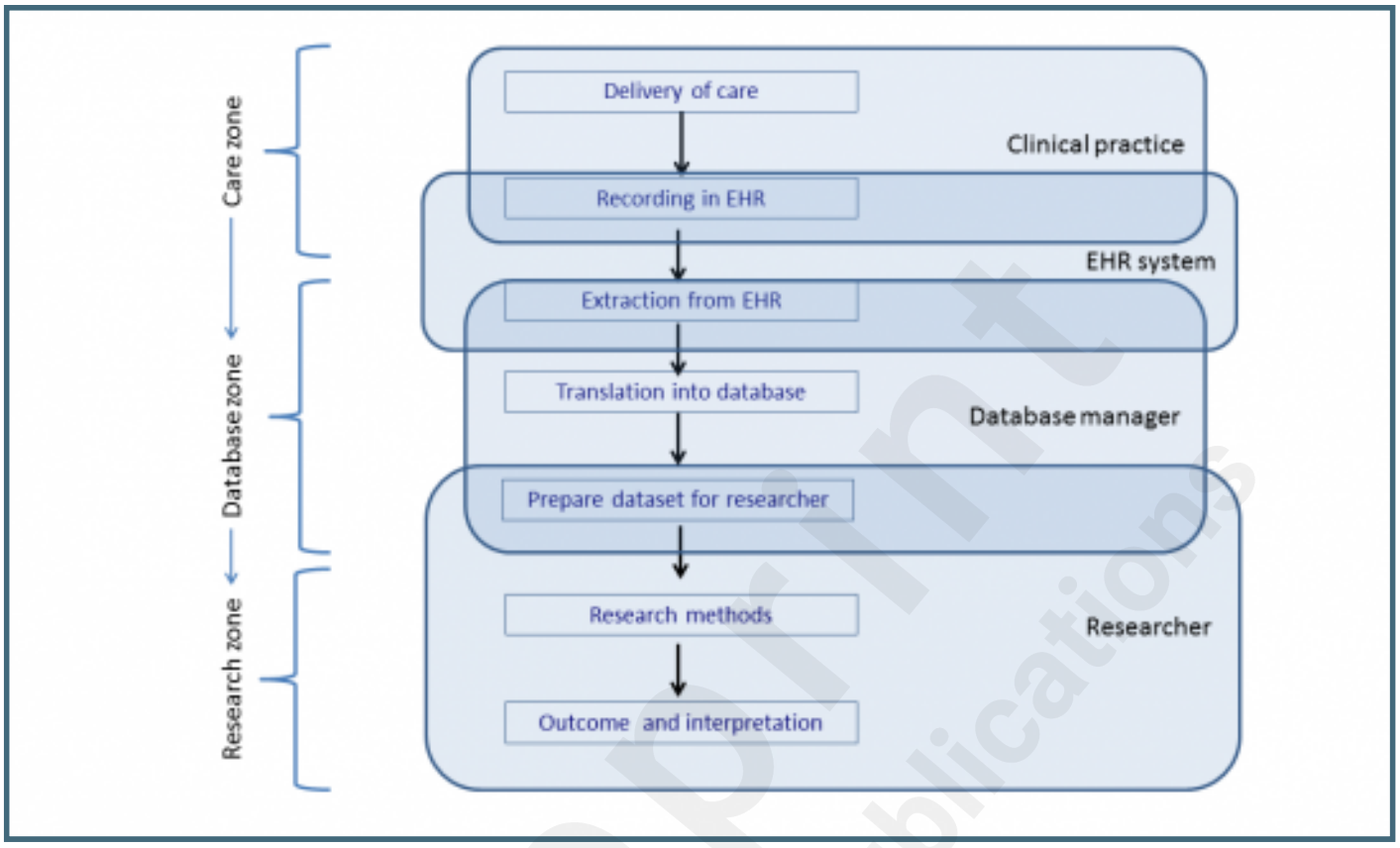
opinion and user experience. BMJ quality & safety; 2021. 30(12): p. 1010-1020. PMID:33963072
19. Månsson, J, et al., Collection and retrieval of structured clinical data from electronic patient records in general practice A first-phase study to create a health care database for research and quality assessment. Scandinavian Journal of Primary Health Care; 2004. 22(1): p. 6-10. PMID:15119513
20. Arslan, IG, et al., Estimating incidence and prevalence of hip osteoarthritis using electronic health records: a population-based cohort study. Osteoarthritis and Cartilage; 2022. 30(6): p. 843-851. PMID:35307534
21. van den Dungen, C, et al., Do practice characteristics explain differences in morbidity estimates between electronic health record based general practice registration networks? BMC Family Practice; 2014. 15(1): p. 1-7. PMID:25358247
22. Haneuse, S and Daniels, M, A general framework for considering selection bias in EHR-based studies: what data are observed and why? EGEMs; 2016. 4(1). PMID:27668265
23. Grobbee, DE, et al., The Utrecht Health Project: optimization of routine healthcare data for research. European journal of epidemiology; 2005. (20): p. 285-290. PMID:15921047
24. Homburg, MT, et al., Dutch GP healthcare consumption in COVID-19 heterogeneous regions: an interregional time-series approach in 2020-2021. BJGP open; 2023. PMID:38128964
25. Bos, I, et al., Comparison of methods to identify and characterize Post-COVID syndrome using electronic health records and questionnaires. Research Square Company; 2023. DOI:10.21203/rs.3.rs-3255500/v1
26. Blanker, MH, General practice Research Infrastructure Pandemic Preparedness Program (GRIP3). COVID-19 2021. ZonMW project, 2023.
27. Homburg, M, et al., A Natural Language Processing Model for COVID-19 Detection Based on Dutch General Practice Electronic Health Records by Using Bidirectional Encoder Representations From Transformers: Development and Validation Study. Journal of Medical Internet Research; 2023. (25): p. e49944. PMID:37792444
28. Twickler, R, et al., Data resource profile: registry of electronic health records of general practices in the north of the Netherlands (AHON). International Journal of Epidemiology; 2024. 53(2). DOI:10.1093/ije/dyae021
29. Nivel, Nivel Primary Care Database. Nivel website. 2024.
30. Soler, JK, et al., The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care. Family practice; 2008. 24(4). PMID: 18562335
31. Nahler, G, et al., Anatomical therapeutic chemical classification system (ATC). Dictionary of pharmaceutical medicine; 2009. 8. DOI: 10.1007/978-3-211-89836-9_64
32. Westerdijk, M, et al., Defining care products to finance health care in the Netherlands. The European Journal of Health Economics; 2012. 13: p.203-21. PMID: 21350859
33. Nielen, MM, et al., Estimating morbidity rates based on routine electronic health records in primary care: observational study. JMIR medical informatics; 2019. 7(3): p. e11929. PMID:31350839
34. Hek, K, et al., Antibiotic prescribing in Dutch daytime and out-of-hours general practice during the COVID-19 pandemic: a retrospective database study. Antibiotics; 2022. 11(3): p. 309. PMID:35326772
35. Zorginstituut Nederland, Farmacotherapeutisch Kompas. Farmacotherapeutisch Kompas website. 2024.
36. Wilkinson, MD, et al., The FAIR Guiding Principles for scientific data management and stewardship. Scientific data; 2016. 3(1): p1-9. PMID: 26978244
37. Jacobsen, A, et al., A generic workflow for the data FAIRification proces. Data

Intelligence; 2020. 2(1-2): p56-65. DOI:10.1162/dint_a_00028
38. Chang, E, et al., The use of SNOMED CT, 2013-2020: a literature review. Journal of the American Medical Informatics Association; 2021. 28(9). PMID: 34151978
39. Heins, MJ, et al., Opportunities and obstacles in linking large health care registries: the primary secondary cancer care registry-breast cancer. BMC medical research methodology; 2022. 22(1): p. 124. PMID:35477392
40. Liaw, S-T, et al. Data quality and fitness for purpose of routinely collected data–a general practice case study from an electronic practice-based research network (ePBRN). in AMIA Annual Symposium Proceedings; 2011. American Medical Informatics Association. PMID:22195136
41. Neiva, FW, et al., Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature. Information and Software Technology; 2016. (72): p. 137-150. DOI:10.1016/j.infsof.2015.12.013
42. Gini, R, et al., Data extraction and management in networks of observational health care databases for scientific research: a comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE strategies. Egems; 2016. 4(1). PMID:27014709
43. Kahn, MG, et al., A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems; 2016. 4(1). PMID:27713905
44. Blacketer, C, et al., Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association; 2021. 28(10): p. 2251-2257. PMID:34313749
45. Dros, JT, et al., Detection of primary Sjögren's syndrome in primary care: developing a classification model with the use of routine healthcare data and machine learning. BMC Primary Care; 2022. 23(1): p. 199. PMID:35945489

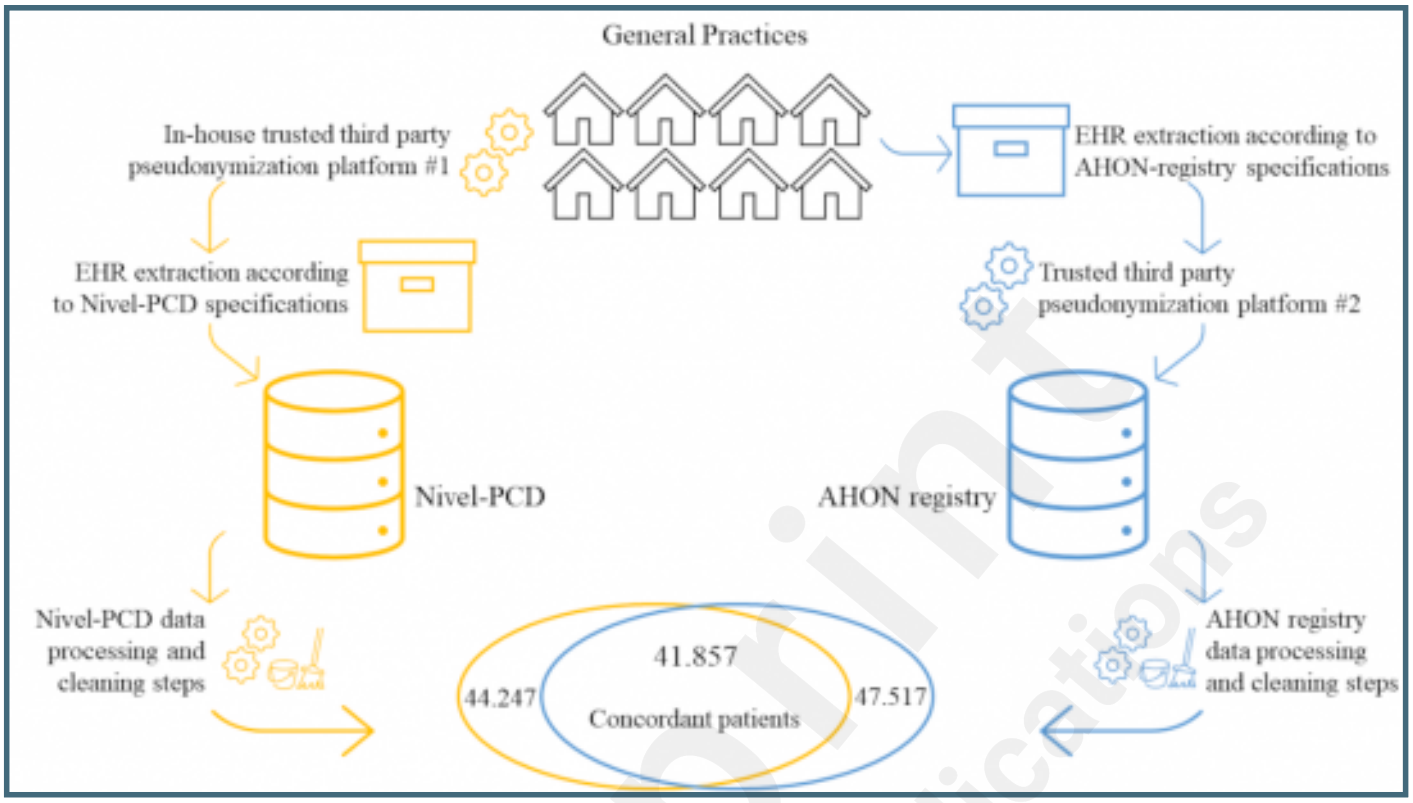
Supplementary Files

Figures

Steps and actors involved in the data flow between the delivery of care and applications reusing the data, Verheij et al. [1].



Flowchart of patients included in the study population.



Schematic overview of the variables and their relationship and origin.