METHOD ARTICLE

Extraction of CPRD additional clinical data using R [version 1; peer review: 2 approved with reservations]

Anthony Nash , M. Zameel Cader

Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, UK

V1 First published: 11 Sep 2020, 9:1124
<https://doi.org/10.12688/f1000research.26228.1>
Latest published: 11 Sep 2020, 9:1124
<https://doi.org/10.12688/f1000research.26228.1>

Abstract

The Clinical Practice Research Datalink is a nation-wide database of primary healthcare data records in England (UK) linked to several health services. A visit to a health practitioner can result in the digital storing of diagnostic and prescription therapeutic information. Access to patient primary care and linked service data depends on the research in mind; however, typically several flat files that describe patient interactions with a health practitioner are delivered. Some of these files will describe additional data such as the result of medical tests and patient lifestyles, denoted collectively into entity values. This data is used to supplement the medical notes recorded by a general practitioner. We have made available a set of R scripts that reads the clinical flat files, additional clinical flat files and entity values, and returns patient clinical data linked with the requested additional data. We have also included *medcode* descriptions associated with several entities along with instruction of how to extend the code for additional entities. The code is free to download under the MIT license: https://github.com/acnash/CPRD_Additional_Clinical

Keywords

CPRD, Primary Care, Electronic Healthcare Records, Epidemiology, R

Open Peer Review

Approval Status ? ?

	1	2
version 1		
11 Sep 2020	view	view
1. Zhi Yang , NanoString Technologies, Inc., Los Angeles, USA		
2. Jessica Pinaire , Montpellier University, Montpellier, France		
Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **RPackage** gateway.

Corresponding author: Anthony Nash (anthony.nash@ndcn.ox.ac.uk)

Author roles: **Nash A:** Conceptualization, Data Curation, Methodology, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Cader MZ:** Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: National Institute for Health Research Oxford Health Biomedical Research Centre (grant BRC-1215-20005). The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health. The script was used for part of the ISAC project 17_255R2.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Nash A and Cader MZ. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Nash A and Cader MZ. **Extraction of CPRD additional clinical data using R [version 1; peer review: 2 approved with reservations]** F1000Research 2020, 9:1124 <https://doi.org/10.12688/f1000research.26228.1>

First published: 11 Sep 2020, 9:1124 <https://doi.org/10.12688/f1000research.26228.1>

Introduction

The Clinical Practice Research Datalink (CPRD) is an NHS primary care service that stores clinical, referral, therapeutic data and linked medical services such as imaging and hospital records of patients enrolled in England (UK). In its current form, CPRD has been active for over twenty years and holds up to 11 million records¹. CPRD data has been widely used to improve medical practice and to further our understanding of drug efficacy and drug safety and disease mechanism². As with all longitudinal data, understanding the outcome of a disease is usually confounded by several factors³. Patient comorbidities, therapy data and social habits (to name but a few) are all potential confounders that should be treated appropriately (see 4 for a research example of common patient characteristics using a CPRD data release). However, the curation and manipulation of CPRD data can cause significant barriers for many researchers, especially those less familiar with manipulating text-based large datasets or programming.

Data released from the CPRD Gold database data is presented as a series of flat files. A patient's longitudinal data is presented across several rows, with each column denoting a record property. Patient records are linked between flat files using an anonymized patient identifier and several additional identifiers depending on the data at hand. Clinical diagnosis and prescribed therapeutics is encoded as a *medcode* and a *prodcodes*, with a corresponding description available using the CPRD data dictionary. Most requests for CPRD Gold primary care data will usually return several clinical flat files. These are primary care records added by the patient's General Practitioner (GP). The GP may also record additional clinical information, such as whether the patient smokes, how much they weigh, how often they drink, or the results from medical tests. This data is contained in a linked additional clinical set of flat files.

We present an R script that we have used to retrieve additional clinical data for patients. The script contains the code necessary to retrieve several patient characteristics (smoking, alcohol consumption, etc.) and is relatively straightforward to extend. We aim to continue updating our GitHub release with further additional clinical properties and any corresponding *medcode* descriptions whilst our projects remain active. In this manuscript, we outline the link between clinical and additional clinical patient data, the execution flow of the R scripts, and how to expand on the existing code. The R script is available for free and to download under an MIT license on GitHub: https://github.com/acnash/CPRD_Additional_Clinical

Methods

We describe the link between patient clinical data and additional clinical data in cases where such a CPRD data release has been made available.

Both clinical data and additional clinical data are stored in two separate sets of files, often denoted as *head_Extract_Clinical_###.txt* and *head_Extract_Additional_###.txt*, respectively (this may depend on the service provided by those who extract the CPRD data). Additional clinical data, stored in the second set of files, is typically accompanied by a lookup data table with lookup codes stored in text files. Additional clinical data is identified by a unique entity (*enttype*, the *entity type*) value, for example, smoking status has an entity value of 4. Both clinical files and additional clinical files will have an *enttype* column. A researcher can retrieve all patients with a smoking status using this column. Then, depending on the particular additional information of interest, data on that entity can be retrieved from up to seven data fields, for example, whether they smoke, how many packs of cigarettes per week, whether they smoke a pipe, and how many ounces of tobacco. Some values are encoded and require the corresponding lookup files which share the same names as the entry in the data lookup column. For example, smoking status is defined in the first data column and the values are found in the YND.txt lookup file as denoted in the corresponding first data lookup column.

We have built an extensible R script which parses the clinical and additional clinical text files for an entity and then returns all clinical data with the corresponding additional clinical data for those patients with an entity of interest on record. As this is longitudinal data, patients may have several additional clinical rows on record for one entity. Therefore, the *adid* value (present in both sets of records) is used to match the clinical data with additional clinical data. If there are additional clinical records without a matching clinical record, the code ignores and moves to the next patient. We have also added *medcode* descriptions for those entities we are currently using. As our research expands so too will this list.

The R code is easy to expand and allows the researcher to treat each entity according to their needs. As the code stands, it currently supports our active research and we have been able to retrieve the additional clinical values and paired clinical *medcode* descriptions for smoking, alcohol consumption, exercise, and weight/BMI. A user need

only execute several short lines of R code to retrieve any of the entities. For example, to retrieve patient data on smoking habits:

```
#source the R core
source("CPRDLookups.R")

#define the location of both sets of files
additionalFiles <- "./head_Extract_Additional_folder"
clinicalFiles <- "./Clinical_folder"

#add the file locations to a list with element names
additionalFileList <- list(additional=additionalFiles, clinical=clinicalFiles)

#look up smoking additional data
resultDF <- getEntityValue("smoking", additionalFileList, idList)
```

The *idList* is an R list or vector of *patids* (patient identifiers) of those patients of interest. The *additionalFileList* list contains the dedicated location of both sets of files. The content of these files is loaded into data frames and stored in the R environment. Ensure that there is enough system RAM available to open CPRD data files.

The entity value *smoking* is used to instruct the algorithm to look for all clinical and additional clinical data with an entity value of 4. A list of currently supported entities can be retrieved by executing *getEntityValue()*. All matching clinical data with additional clinical data (matched using the *adid* identifier if there are multiple rows per patient) are returned as a data frame with an additional column for *medcode* description. The *medcode* descriptions have been added into the R script and are used by the entity values currently supported.

The user executes one function and the existing R framework steps through a series of predefined functions before entering the user's bespoke function (Figure 1). To extend the code to support additional clinical data, the researcher need only expand the *if()* statement in the function *getEntityValue()* with a function call to their own *get<enttype>Data()* function, following the format presented in the functions *getExerciseData()*, *getSmokingData()*, etc. Whether the researcher wishes to expand the results with a description of the medcodes is controlled by calling the defined function *addMedcodeDescription(resultDF, <medcode_description_list>)*

Figure 1. The R code-flow demonstrating the steps from the user-facing *getEntityValue()* through to loading the CPRD data files; parsing for entity values for “smoking”, and adding *medcode* descriptions. The function *getSmokingData()* in the orange square can be replaced by a user defined function with the same function arguments for other entities.

within the *if()* statement. Both areas are illustrated below, with the potential code extension commented and a user defined entity name denoted with the place holder *new_entity*:

```
if(tolower(entityString)=="smoking") {
  resultDF <- getSmokingData(idList, additionalClinicalDataDF, clinicalDataDF)
  resultDF <- addMedcodeDescription(resultDF, smokingMedcodeDescriptionList)
} else if(tolower(entityString) == "bmi") {
  resultDF <- getBMIData(idList, additionalClinicalDataDF, clinicalDataDF)
  resultDF <- addMedcodeDescription(resultDF, bmiMedcodeDescriptionList)
}
#####
##Add more entity statements here....
#if(tolower(entityString)=="new_entity") {
#  resultDF <- getnew_entityData(idList, additionalClinicalDataDF, clinicalDataDF)
#  resultDF <- addMedcodeDescription(resultDF, new_entityMedcodeDescriptionList)
#}
#####
else {
  print("Unrecognised entity type. Try one of the following:")
  outputCurrentOutput()
  return(NULL)
}
```

Unfortunately, it is not entirely possible to automate the decoding of the additional clinical data for every entity. There currently 501 unique entities (CPRD Gold release February 2018) and 91 unique lookup text files. As presented, we have used this code to return the decoded entity data along with a corresponding clinical *medcode* description. However, depending on the research, a user can return a result bespoke to their needs by simply defining a new function. Up to that stage, the existing code finds and retrieves all data with a matching *entity* value.

The script requires the R data.table package.

Results

We present a sample of a use-case where clinical records were linked with smoking patient data. The data presented in this manuscript and uploaded to the GitHub repository has been fabricated in the interest of patient confidentiality. A test data set of 2,000 clinical primary care records were parsed for additional data concerning smoking. Of those patients, 1,993 had a record for smoking. Having matched those patients, retrieved the data and decoded their smoking status for either "Data not entered", "Yes", "No" or "Ex smoker", each *medcode* from a corresponding clinical record was populated with a readable description. A snapshot of fabricated patient records is presented in [Figure 2](#).

	patid	eventdate	sysdate	constype	consid	medcode	staffid	episode	enttype	adid	Smoking	Cigs_per_day
2593	1245	1994-03-11	1994-05-19	2	5182	2111	1007	0	4	4776	Yes	8
2595	289	1996-08-21	1996-10-03	2	2490	1823	5630	0	4	2735	Yes	
2611	2363	2009-05-07	2009-06-25	2	3441	90	1851	0	4	723	Ex smoker	
2612	2843	2014-11-09	2014-12-13	2	1089	93	5702	0	4	3956	Yes	4
2613	4825	2014-06-01	2014-08-14	2	2059	90	4468	0	4	4976	Ex smoker	
2969	4163	2007-09-22	2007-10-17	2	5317	90	2266	0	4	4239	Ex smoker	
2974	5465	2009-11-08	2009-12-01	2	4248	90	711	0	4	1031	Ex smoker	
2977	236	2012-02-29	2012-05-12	2	615	90	174	0	4	442	Ex smoker	
2980	791	2013-12-20	2014-01-10	2	8	90	4366	0	4	5702	Ex smoker	
5396	1005	1994-03-16	1994-04-03	2	1056	60	3597	0	4	1703	No	
	Cigars_per_day ounces_tobacco_per_day startdate stopdate data7 medcode_description											
2593										<NA>	Health ed. - smoking	
2595										<NA>	Smoker	
2611										<NA>	Ex smoker	
2612										<NA>	Cigarette smoker	
2613						01/01/2014				<NA>	Ex smoker	
2969										<NA>	Ex smoker	
2974										<NA>	Ex smoker	
2977						01/11/1995				<NA>	Ex smoker	
2980										<NA>	Ex smoker	
5396										<NA>		

Figure 2. A snapshot of the ten rows of patient data showing smoking status along with additional smoking data columns, the corresponding primary care *medcode* and the data (*eventdate*) in which the information was recorded by the GP. All dates and data have been fabricated.

Conclusion

We present an R script that searches for and retrieves additional clinical entity values and combines the results with patient clinical data. Although access to CPRD data requires an entrusted license holder, not all institutes will be fortunate enough to have access to software necessary to perform CPRD data curation and analysis⁵. At the time of writing, the software can support smoking, alcohol consumption, exercise and weight/BMI entity lookups. The script has been designed in such a way that most R users can extend the code and provide unique behaviour for a lookup entity of interest. We will continue to support the scripts, adding in new entities and *medcode* descriptions as our research requires. This script is part of a family of R packages and scripts currently under development.

Data availability

Zenodo: CPRD_Additional_Clinical, <http://doi.org/10.5281/zenodo.3989634>⁶.

This project contains the following underlying data:

- smokingDF.rda - a fabricated result R data frame, as presented in the Results section.

Software availability

R script available from: https://github.com/acnash/CPRD_Additional_Clinical

Archived R script as at time of publication: <http://doi.org/10.5281/zenodo.3989634>⁶.

License: MIT

RRID:SCR_018959

The script also holds *medcode* descriptions taken from the CPRD data dictionary for several entities. Samples of the clinical data and additional clinical data have not been provided as access to this data is only permitted following execution of the appropriate license agreements. If access to any of this data is required, we can provide *medcode* lists and the criteria used to extract the dataset (ISAC project 17_255R2) used to create this software, and we recommend contacting CPRD directly.

References

1. Herrett E, Gallagher AM, Bhaskaran K, *et al.*: **Data Resource Profile: Clinical Practice Research Datalink (CPRD)**. *Int J Epidemiol.* 2015; **44**(3): 827–836.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Ghosh RE, Crellin E, Beatty S, *et al.*: **How Clinical Practice Research Datalink data are used to support pharmacovigilance**. *Ther Adv Drug Saf.* 2019; **10**: 2042098619854010.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Skelly AC, Dettori JR, Brodt ED: **Assessing bias: the importance of considering confounding**. *Evid Based Spine Care J.* 2012; **3**(1): 9–12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Wright AK, Welsh P, Gill JMR, *et al.*: **Age-, sex- and ethnicity-related differences in body weight, blood pressure, HbA_{1c} and lipid levels at the diagnosis of type 2 diabetes relative to people without diabetes**. *Diabetologia.* 2020; **63**(8): 1542–1553.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Denaxas SC, George J, Herrett E, *et al.*: **Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)**. *Int J Epidemiol.* 2012; **41**(6): 1625–1638.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Nash A: **acnash/CPRD_Additional_Clinical: Initial release in line with F1000 submission and Zenodo link. (Version v0.1)**. *Zenodo.* 2020.
<http://www.doi.org/10.5281/zenodo.3989634>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 18 July 2022

<https://doi.org/10.5256/f1000research.28946.r142507>

© 2022 Pinaire J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jessica Pinaire

Montpellier University, Montpellier, France

The article presents a set of scripts, developed in R, to explore clinical data from CRPD for a given patient.

This data is a big database of primary care in the UK, it contains various information such as a visit to a general practitioner, drug prescriptions, etc. Such a database is a complex database, hard to manage for the researchers not aware of its structure. To bridge the gap between clinical knowledge and data manipulation, the authors suggest a solution through their scripts.

The article is easy to follow and well written.

The article is interesting and highlights the difficulties faced by many biomedical researchers in various fields when they have to use databases with heterogeneous data and complex frameworks. This requires skills in various areas which they do not always have. However, since there isn't a specific method - like in data analysis - with results to present, I wonder if that kind of structure is adapted for this topic instead of a demonstration article.

I have some remarks:

Introduction part:

- Where can we find this data CRPD and who can access it?
- How many researchers use it?
- The article does not present a state of the art on the subject, are there equivalent works for other types of databases?

Method part:

- The author refers to CRPD data throughout the article, but we don't see the data. As there are confidential data, it might be nice to provide a dummy dataset to be referred to throughout the article to make it easier to follow.
- Without data it is impossible to check if the scripts work, in this case, the dummy dataset could

be used to test the different scripts.

- "Ensure that there is enough system RAM available to open CPRD data files." What is the minimum RAM resource needed? Can we do everything? What is the limit?
- "Unfortunately, it is not entirely possible to automate the decoding of the additional clinical data for every entity." Can you explain more? What are the obstacles?
- "There currently 501 unique entities (CPRD Gold release February 2018) and 91 unique lookup text files." I think the verb is missing...

Conclusion part:

- What are the limitations of this work?
- What are the prospects for this work?

Perhaps, it would be better to plan to put all the scripts in a package that is in the usual use of R.

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Health Data, Machine Learning, Text Mining

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 12 May 2021

<https://doi.org/10.5256/f1000research.28946.r84116>

© 2021 Yang Z. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Zhi Yang

NanoString Technologies, Inc., Los Angeles, CA, USA

The author described an R script to quickly extract patient clinical data from CPRD for given patients. As a result, this R script is easy to use and extend based on users' interests. In general, the description is clear and straightforward to follow. However, due to errors in loading the test datasets, it is infeasible to validate whether the R script works as expected fully. Hopefully, the authors can address that issue soon.

1. I tried to download the code from GitHub, and it seems that the `smokingDF.rda` is corrupted. The error message is shown as "bad restore file magic number (file may be corrupted) -- no data loaded". This error remained after trying with different R versions (3.6 and 4.0) and different operating systems (Windows and OS).
2. In the README file, since users don't have access to the example files in Line 36 & 37, it is unlikely to reproduce the same results and examine whether the script works as expected or not. Is it possible to provide some mock datasets for both *additionalFiles* and *clinicalFiles* to test the script?
3. The following comments are specific to the R script without actually running it. I should be able to provide more concrete suggestions after having access to the test datasets.
 - 3a. the *as.data.table* function needs to be added to `@importFrom`. From Line 283 to Line 291 of the R script, please explain why those variables are hard-coded.
 - 3b. In the R script, some documentations are missing for `Title`, `@param`, `@return`, etc., for example, `getEthnicityData`.
 - 3c. In the R script, *idList* is not used in the *getEthnicityData* function.
 - 3d. When generating the file names (Line 327), please use functions like *fs::path* rather than *paste0* with `"\"` which might throw an error at a different operating system.
4. This is just a suggestion for improvement. But it is highly recommended to make an R package rather than providing a single R script. It is easy for authors to maintain and for users to use and navigate. Especially, it is convenient for R developers to contribute in a consistent way like other R packages. To adapt the single R script into the R package structure, the main R script can be stored in the `/R` folder. All the medcode from Line 9 to 278 can be saved under the `/inst` folder as internal data. The example script can be provided as a vignette.
5. The authors mentioned about few scripts are available for researchers to obtain the requested information from CPRD. How about another R package called `rEHR` (<https://github.com/rOpenHealth/rEHR>)? Are there any additional features provided by this R script described here but not by `rEHR`?

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, Bayesian, Machine learning.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research