An Algorithm to Identify Duplicate Patients When Pooling Aggregate Data From Two Primary Care Databases in the United Kingdom

Estel Plana,[1] Leah J. McGrath,[2] Joan Fortuny,[1] Ana Ruigomez,[3] Oscar Fernandez-Cantero,[3] Luis-Alberto Garcia-Rodriguez,[3] Alicia W. Gilsenan,[2] Elizabeth Andrews,[2] Cristina Rebordosa,[1] Ryan Ziemiecki[1]

[1]RTI Health Solutions, Barcelona, Spain; [2]RTI Health Solutions, Research Triangle Park, NC, United States;
[3]CEIFE - Spanish Centre for Pharmacoepidemiologic Research, Madrid, Spain

CONFLICT OF INTEREST

E. Plana, L. McGrath, J. Fortuny, A. Gilsenan, E. Andrews, C. Rebordosa, and R. Ziemiecki are full-time employees of RTI Health Solutions, which received funding from Shire Development LLC to conduct this study. RTI conducts work for government, public, and private organizations, including pharmaceutical companies. A. Ruigomez, O. Fernandez-Cantero, and L. Garcia-Rodriguez are employees at the Spanish Centre for Pharmacoepidemiologic Research, which collaborates with pharmaceutical companies, regulatory authorities, and contract research organizations. Authors had full access to the study data and had final responsibility for the development, finalization, submission, and presentation of the poster.

ABSTRACT

Background: The Clinical Practice Research Datalink (CPRD) and The Health Improvement Network (THIN) are two similarly structured, deidentified electronic medical record databases in the United Kingdom. To increase the number of patients available, both data sources can be pooled. However, some practices provide data to both databases, and duplicate patients should be identified and steps taken to avoid double-counting patients and study outcomes.

Objectives: To describe a patient-level algorithm to deduplicate patients in CPRD and THIN using a cohort of prucalopride users.

Methods: Adult users of prucalopride were identified in CPRD and THIN, from April 2010 through May 2014, in England, Wales, and Northern Ireland. Patients were considered duplicated if they had the same value for year of birth, sex, region, month and year of at least one prucalopride prescription, and either the same registration date or family ID. For potentially duplicated patients with a discrepancy in the number of prescriptions, all drugs prescribed during the study period were manually reviewed. A practice was considered duplicated in CPRD and THIN if at least one patient was found to be duplicated. Duplicate practices were retained in CPRD if the practice participated in linkage with the national death register at the Office for National Statistics (ONS) and Hospital Episode Statistics (HES), otherwise the practice was retained in THIN.

Results: There were 994 users of prucalopride in CPRD and 808 in THIN. The deduplication algorithm identified 424 duplicate patients. Manual review of an additional 95 potentially duplicate patients with discrepant prescriptions identified 86 additional duplicate patients. There were 214 duplicate practices. Pooling the databases increased the number of available prucalopride users by 30% had only CPRD been used and by 60% had only THIN been used.

Conclusions: Pooling of data from similar databases is a convenient way to increase study size. Using patient-level demographics and pharmacy data can identify duplicate patients and practices, allowing reliable deduplication in CPRD and THIN without compromising patient or practice confidentiality.

BACKGROUND

- To maximize the study size for a multidatabase study evaluating the cardiovascular safety of prucalopride (EU PAS Register Number: EUPAS9200), data from the CPRD and THIN were combined.
- The CPRD and THIN include deidentified records from general practitioners in the United Kingdom that are made available to researchers.
- Table 1 shows the types of information available in each data source.
- There is a certain number of practices that contribute data to both the CPRD and THIN, making it necessary to deduplicate them when combining data from both data sources in order to avoid double counting of study subjects.
- Because data are deidentified prior to releasing for research, an algorithm must be used to identify and remove overlapping patients.

Table 1. Comparison of Data Sources

Characteristic	CPRD	THIN
Population covered	4.4 million[a]	3.6 million[b]
Number of practices	674	587
Linkage to HES data	~50%	~30%
Linkage to ONS data	~50%	No
Free text available	No	Yes
GP questionnaire possible	Yes, for active practices	Yes, for approximately 50%-60% of practices

[a] Herrett et al. (2015).[1]
[b] IMS Health (2015).[2]

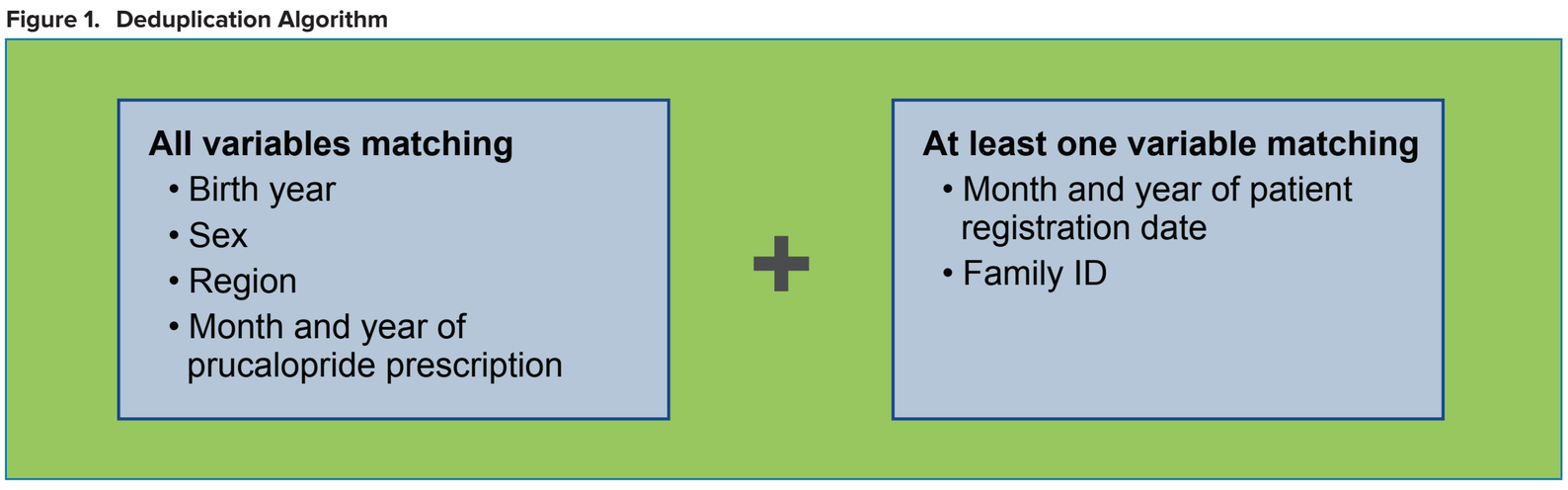
OBJECTIVE

- To describe a patient-level algorithm to identify and remove duplicate practices and patients in the CPRD and THIN using a cohort of prucalopride users.

METHODS

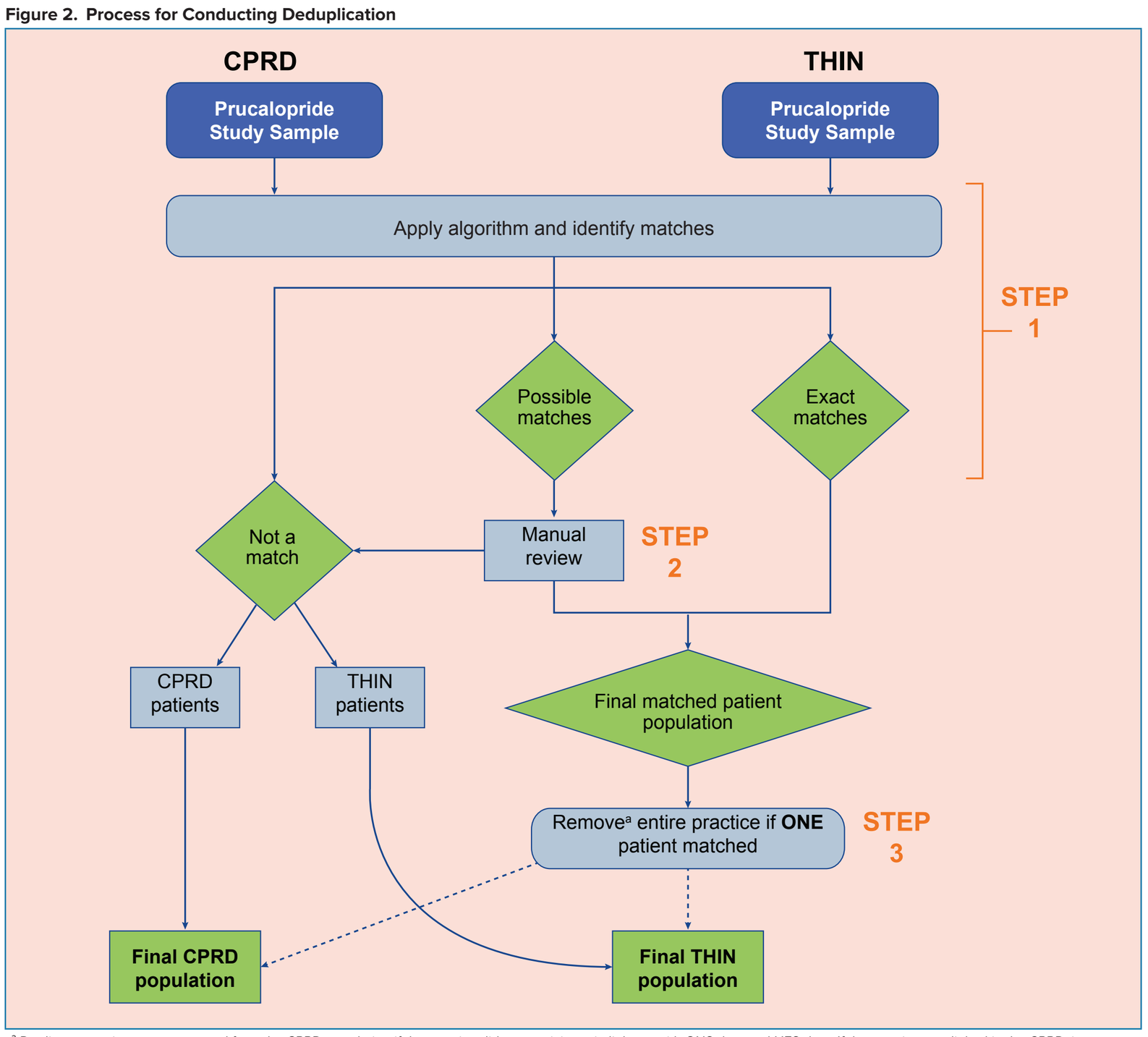
- The study population comprised adult patients with a prescription for prucalopride in the CPRD and THIN from April 2010 through November 2014 (CPRD) or September 2014 (THIN) who lived in England, Wales, or Northern Ireland [Note: Additional data were available since the time the abstract was submitted].
- Figure 1 shows a description of the deduplication algorithm. The algorithm was developed based on the method described by Cai et al. (2012)[3] but adapted to be applied at the patient level. The algorithm compares patient-level variables in each database to classify patients as matches (i.e., potential duplicates).

Figure 1. Deduplication Algorithm



- A three-step process was applied to identify and remove duplicate patients (Figure 2).
- **Step 1:** The algorithm was applied to both databases, and patients were classified as (1) not a match, (2) exact match, or (3) possible match.
  - A match was defined as two patients who were found to share the same value for year of birth, sex, region, month, and year of at least one prucalopride prescription, and either the same registration date or family ID.
    - An **exact match** was defined as two patients who matched on the algorithm criteria and had no more than one discrepancy in the number of prucalopride prescriptions in the electronic medical record.
    - A **possible match** was defined as two patients who matched on the algorithm criteria, but had more than one discrepancy in the number of prucalopride prescriptions in the electronic medical record.
  - All other situations were considered as **not a match**.
- **Step 2:** Possible matches were reviewed manually. All drugs prescribed during the study period were reviewed to determine whether there were identical dates for some of the prescriptions, which was suggestive that they were the same patient.
- **Step 3:** Duplicate practices were removed from one of the databases.
  - A practice was considered to be duplicated in the CPRD and THIN if at least one duplicate patient record was found in each database.
  - To maximize supplemental data available for the overarching study, duplicate practices were retained in the CPRD if the practice participated in the linkage with ONS data and HES data; otherwise, the practice was retained in THIN, which had access to free text information. For three linked practices from the CPRD that did not accept questionnaires, the corresponding practices in THIN were retained.

Figure 2. Process for Conducting Deduplication

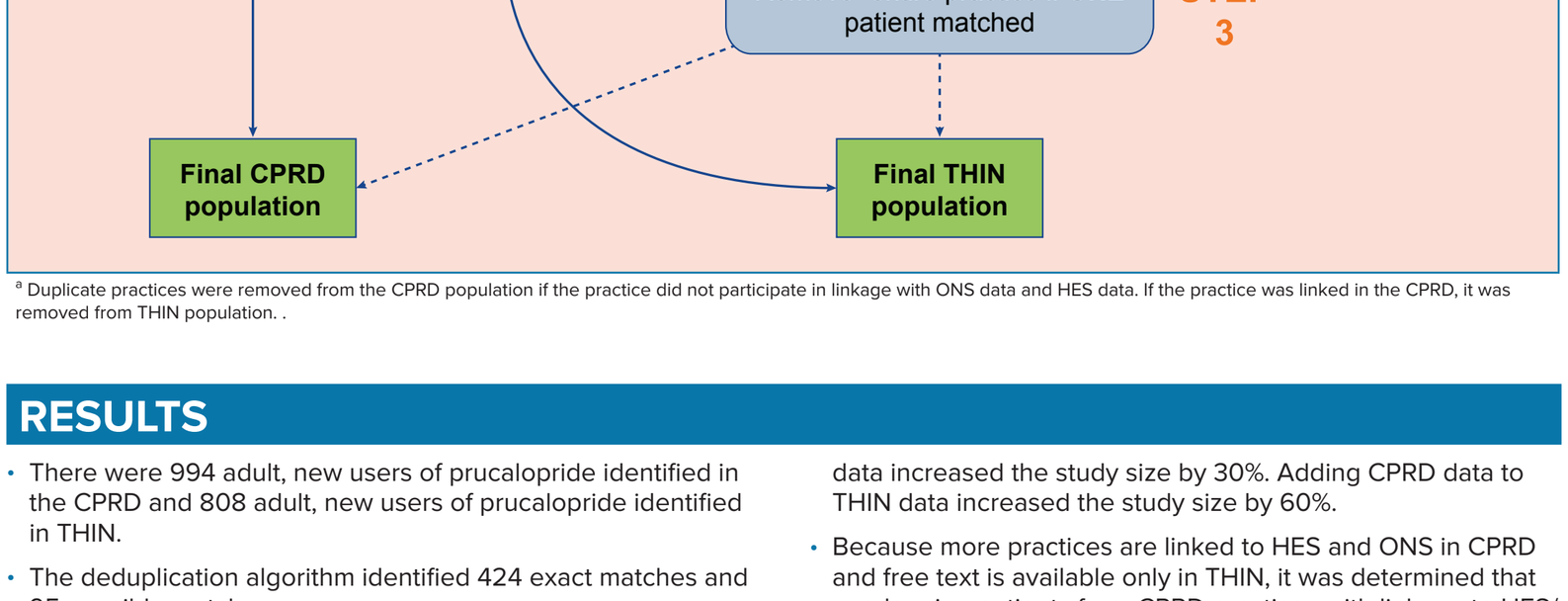


[a] Duplicate practices were removed from the CPRD population if the practice did not participate in linkage with ONS data and HES data. If the practice was linked in the CPRD, it was removed from THIN population. .

RESULTS

- There were 994 adult, new users of prucalopride identified in the CPRD and 808 adult, new users of prucalopride identified in THIN.
- The deduplication algorithm identified 424 exact matches and 95 possible matches.
- Out of the 95 possible matches with discrepant prescriptions, manual review classified 86 of these patients as matches, resulting in 510 overlapping patients receiving prucalopride in CPRD and THIN from 214 different practices.
- As shown in Figure 3, prior to applying study exclusion criteria and selecting the study cohort, adding THIN data to CPRD data increased the study size by 30%. Adding CPRD data to THIN data increased the study size by 60%.
- Because more practices are linked to HES and ONS in CPRD and free text is available only in THIN, it was determined that overlapping patients from CPRD practices with linkage to HES/ONS be retained for the study cohort; otherwise, THIN patients would be retained.
- After removing duplicate practices based on the condition of availability to link to HES and ONS and applying study exclusion criteria, the total sample size was reduced to 1,037 with 600 prucalopride-initiators in the CPRD and 437 prucalopride-initiators in THIN.

Figure 3. Results of Applying the Algorithm to the Cohort of Patients Initiating Prucalopride in the CPRD and THIN



CONCLUSIONS

- Pooling of data from similar databases is a useful way to increase study size.
- The analysis of patient-level demographics and prescription data can be used to identify duplicate patients and duplicate practices in the CPRD and THIN without compromising patient or practice confidentiality.

REFERENCES

1. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015 Jun;44(3):827-36.
2. IMS Health; 2015. Available at: http://www.epic-uk.org/. Accessed June 14, 2016.
3. Cai B, Xu W, Bortnichak E, Watson DJ. An algorithm to identify medical practices common to both the General Practice Research Database and The Health Improvement Network database. Pharmacoepidemiol Drug Saf. 2012 Jul;21(7):770-4.

CONTACT INFORMATION

Estel Plana, MSc
Senior Statistician

RTI Health Solutions
Trav. Gracia 56, Atico 1
08006 Barcelona, Spain

Phone: +34.93.362.2832
E-mail: eplana@rti.org

Presented at: 32nd International Conference on Pharmacoepidemiology & Therapeutic Risk Management, August 25-28, 2016, Dublin, Ireland