# A review of routinely collected data studies in urology: Methodological considerations, reporting quality, and future directions



*Blayne Welk, MD[1]; Justin Kwong, MD[2]*

[1]Department of Surgery and Epidemiology and Biostatistics; [2]Department of Surgery; Western University, London ON, Canada

Cite as: *Can Urol Assoc J* 2017;11(3-4):136-41. http://dx.doi.org/10.5489/cuaj.4101

## Abstract

Studies using routinely collected data (RCD) are common in the urological literature; however, there are important considerations in the creation and review of RCD discoveries. A recent reporting guideline (REporting of studies Conducted using Observational Routinely-collected health Data, RECORD) was developed to improve the reporting of these studies. This narrative review examines important considerations for RCD studies. To assess the current level of reporting in the urological literature, we reviewed all the original research articles published in *Journal of Urology* and *European Urology* in 2014, and determined the proportion of the RECORD checklist items that were reported for RCD studies. There were 56 RCD studies identified among the 608 articles. When the RECORD items were considered applicable to the specific study, they were reported in 52.5% of cases. Studies most consistently (>80% of them) reported the names of the data sources, the study time frame, the extent to which the authors could access the database source, the patient selection, and discussed missing data. Few studies (<25%) discussed validation of key coding elements, details on data-linkage, data-cleaning, the impact of changing eligibility over time, or provided the complete list of coding elements used to define key study variables. Reporting factors specifically relevant in RCD studies may serve to increase the quality of these studies in the urological literature. With increased technological integration in healthcare and the proliferation of electronic medical records, RCD will continue to be an important source for urological research.

## Introduction

Routinely collected data (RCD, also known as administrative data) is increasingly being used to answer clinical research questions. While there are many sources of observational data (for example, population surveys or patient registries), RCD is specifically defined as data that is routinely collected for a purpose other than research. Common examples of RCD sources include physician billing records, insurance company records, or government-mandated healthcare utilization records.[1] There are several important strengths that have driven the increased use of RCD research over the last two decades. These include low study costs (especially when compared to prospective clinical studies or randomized, controlled trials), rapid delivery of results, substantial sample sizes, improved generalizability, long-term patient followup (often across different hospitals and physician practices), and the ability to identify very rare patient populations or outcomes.

Given the apparent increase in publications using RCD in many urological journals, we sought to review several key issues related to this research area that are relevant for those conducting these studies, and for clinicians who wish to critically read these studies, understand potential methodological limitations, and evaluate the strength of the study's conclusions.

## Methodological considerations for the use of RCD

Any study using RCD to measure an outcome or an association between two groups will have the usual risks of bias common to all observational studies.

### 1. Selection bias

This occurs when the study population is not a random sample from the target population in which the study conclusions will be generalized. The use of RCD (especially data from countries with socialized healthcare systems, such as Canada) helps to mitigate this bias, as all patients within a population are potentially eligible.

### 2. Information bias

This is defined as the inexact measurement of a variable. This is a significant risk in RCD studies, as researchers have

no control over the data elements or data entry, and in many cases, don't have a complete picture of the basic characteristics of a variable. Perhaps the most famous urological example of this was the temporary removal of prostate-specific antigen (PSA) values from the Surveillance, Epidemiology, and End Results (SEER) dataset due to a large number of incorrect values.[2] With RCD, information bias occurs if coding errors are made at random (and therefore biases the association towards the null hypothesis[3]). When the misclassification is different among groups (non-random), the direction of the bias is not predictable, and can have a significant effect on the estimated association.[4,5]

### 3. Confounding

This occurs when the relationship between the exposure of int erest and the outcome is distorted by another variable (the confounder). Known confounders can be controlled for in the design phase by restriction, or in the analysis phase by multivariable regression analysis or stratification.[6] Residual confounding is a consistent limitation of all studies involving RCD. Propensity scores (which allow the creation of two groups of patients with similar overall characteristics across a large number of explicitly defined variables) or high-dimensional propensity scores (unique to RCD) can help address confounding.[7,8] Alternatively, instrumental variables (which result in a quasi-randomization of patients) can be used to address unmeasured confounding.[9]

## Special considerations for the use of RCD

RCD studies also have unique features that should be considered. In clinical research, variables are often directly observed by a physician or extracted from the medical record by a researcher with a specialized knowledge of the relevant study variables. In most cases, this allows for a high degree of confidence in the accuracy of individual variables. In contrast, RCD studies rely on a code as a surrogate for a clinical variable (for example, an ICD code for prostate cancer, or a physician billing code for a radical prostatectomy). The completeness of coding for a particular condition is dependent on the rigour and incentives driving data-collection, and will vary by condition. While the diagnosis of a myocardial infarction is highly likely to be correctly documented, other conditions, such as male lower urinary tract symptoms, are less likely to be documented accurately.[10] Given these limitations, the measurement characteristics of the coding element should be explicitly acknowledged when possible. While some codes have high face validity, ideally key elements of RCD will have been assessed in a validation study to determine precise measurement characteristics (such as sensitivity, specificity, positive and negative predicative values).[11] However, validation studies are often not available or referenced. An assessment of a variety of RCD studies found that only 12% of RCD studies stated or referenced the measurement characteristics associated with key coding elements, and of these, 40% had a probability <50% that the code actually represented the variable of interest.[12]

A large sample size is one of the advantages of RCD; however, this can also make interpretation of the study results and traditional hypothesis-testing challenging. As the sample size increases, a small difference is more likely to become statistical significant. When this difference represents a very clinically important outcome (such as death), then this may still be very relevant. However, in many cases the difference may be clinically insignificant. The use of confidence intervals (in additional to or instead of p values) and reporting the absolute risk difference provides more relevant information to the reader, and is especially pertinent when the sample size is greater than 100 000.[13]

## Reporting quality of RCD studies in urology

Reporting standards have been developed for use with many study designs, such as randomized, controlled trials (CONsolidated Standards Of Reporting Trials, CONSORT)[14] and observational studies (STrengthening the Reporting of OBservational studies in Epidemiology, STROBE).[15] While the STROBE guideline applies to all observational studies, it was recognized that unique characteristics of RCD required the inclusion of additional reporting characteristics. Last year, the RECORD (REporting of studies Conducted using Observational Routinely-collected health Data) statement was published as a supplement to the STROBE statement to specifically address important elements of a RCD study.[16] Thirteen additional complementary items were recommended after an extensive collaborative process.[17] These reporting guidelines are not meant to measure study quality, only to outline a minimum and transparent level of reporting for key study elements.[16,18]

## Current status of reporting of RCD studies in urology

To understand the current level of reporting in RCD studies in urology, we identified all the RCD original research studies published in the *Journal of Urology* (*JU*) and *European Urology* (*EU*) in 2014. RCD studies were specifically selected, therefore, those that used other observational data (such as secondary analysis of prospective cohorts or survey data) were not included. The complete table of contents for each journal issue was reviewed, and where necessary, abstracts or full-text versions reviewed by two physicians to confirm that the study used RCD. We identified relevant studies based on a review of the paper's methods section for common RCD sources (such as Medicare); any data sources that

may have represented RCD were independently assessed for their applicability using online resources.

We identified 56 RCD studies (40 *JU*, 16 *EU*) from among the 608 articles. In *JU*, the most common country of origin of the senior author was the U.S. (87.5%) and the most common subspecialty areas were oncology (52.5%), pediatrics (20.0%), endourology (12.5%), and other (15.0%). In *EU*, the most common country of origin of the senior author was the U.S. (50.0%) and the most common subspecialty area was oncology (93.3%). The most frequently used RCD sources were SEER (15 studies), Medicare sample (nine studies), and the Nationwide Inpatient Sample (five studies). The most common use of multiple RCD sources was a SEER-Medicare linkage (eight studies). The source of the RCD for the reviewed studies was primarily from the U.S. (78.6%, 44/56).

The RECORD checklist[16] was completed for each study, and the reporting of each of the parts of the checklist were classified as present or absent. The RECORD statement has 13 sections, however, for the purposes of assessment of the completeness of reporting, three of them were broken down into separate sections to better assess the specific components (Table 1). Data linkage across multiple RCD sources was not present in 53.6% of the studies, and therefore, this item was considered not applicable for these studies. Overall, when the items were considered applicable, they were reported in 52.5% of cases. Studies (>80% of them) most consistently reported the names of the data sources, the study time frame, the extent to which the authors could access the database source, the patient selection, and discussed missing data. Few studies (<25%) discussed validation of key coding elements, details on data-linkage, data-cleaning, the impact of changing eligibility over time, or provided the complete list of coding elements used to define key study variables.

The importance of these poorly reported elements deserves some additional discussion. Validation studies have already been discussed as an important aspect of a RCD study, particularly for key codes used to define the population, exposure, and outcome. Factors that limit the availability of validation studies are the associated expense and time required (especially when using medical charts as the reference standard), and the fact they are often not seen as a high-impact project by granting agencies or journals. However, by using different linked RCD sources, validation studies can be completed without requiring medical chart abstraction. Data sources such as laboratory values and free text extractions of electronic medical records are novel ways to validate coding elements using existing electronic health data. There are numerous validation studies relevant to urology that have already been published, such as determining the ability of RCD to identify recurrent prostate cancer,[19] upper tract stone disease,[20] adverse urological outcomes of radiation therapy,[21] and other common variables of interest (such as smoking[22] and chronic renal dysfunction).[23] Systematic reviews of validation studies (available for a wide variety of medical diseases) provide an important summary of common conditions, which can be used to make a strong argument about the appropriateness and accuracy of coding schemes.

The publication of the specific codes used to define the RCD variables allows for complete transparency, comparison between studies, and duplication of the methodology by other authors. This also allows those familiar with RCD to assess for the potential for misclassification. Most journals now allow the posting of supplemental information online, which is an appropriate way to document the often extensive coding algorithms used. An alternative is the use of open-source coding repositories (such as ClinicalCodes.org).[24]

Data linkage is often used in RCD studies to increase the available variables, and obtain complementary data on patients over time. Much as the CONSORT flowchart is used in randomized trials, a similar flowchart or description of the number of patients successfully linked, and whether deterministic linkage (based on a definitively unique variable, such as a social security number) or probabilistic linkage (based on a number of variables, such as date of birth, gender, name, region of residence, which when taken together are likely to identify a unique individual) should be considered. Similarly, data-cleaning is an important step in any large dataset to remove values that do not make sense. The extent of the data-cleaning helps communicate the quality of the data and the analytic rigour of the investigators.[25]

Finally, changing eligibility over time is a component of the recommended limitations, which should be addressed in the study discussion. This can occur when there is a shift in the coding structure or coding practice within the RCD source over time. An example is the upcoming shift from ICD-10 to ICD-11 in Canada in the next 5–10 years. If studies include data from both ICD coding schemes, this has to be carefully evaluated to ensure that the coding elements represent the same underlying variable. An easy assessment of this is to look at the coding frequency before and after the change in coding structure; a significant change in the frequency of a code suggests that a change in eligibility occurred. Other reasons for changing eligibility over time in a RCD source could include changes to the data source or emigration.

## Future directions

Electronic data has become a driving force in our society. It has an annual compound growth of 60%, and in 2020 it is estimated there will be 35 zettabytes of electronic data.[26] In healthcare, information technology plays a key role in all aspects of practice — from medical records to medication-prescribing to communication. This wealth of

**Table 1. Proportion of the RECORD items reported in RCD studies from the *Journal of Urology* and *European Urology* in 2014. Each RECORD item is numbered based on the complementary section from the STROBE guidelines**

| RECORD item | Description | Percentage | Proportion |
|---|---|---|---|
| 1.1 | The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. | 83.9% | 47/56 |
| 1.2 | If applicable, the geographic region within which the study took place should be reported in the title or abstract. | 48.2% | 27/56 |
| 1.2 | If applicable, the time frame within which the study took place should be reported in the title or abstract. | 94.6% | 53/56 |
| 1.3 | If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract. | 30.8% | 8/26 (NA for 30 studies) |
| 6.1 | The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. | 60.7% | 34/56 |
| 6.2 | Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. | 21.4% | 12/56 |
| 6.3 | If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage. | 0% | 0/26 (NA for 30 studies) |
| 7.1 | A complete list of codes and algorithms used to classify key variables should be provided. If these cannot be reported, an explanation should be provided. | | |
| | a) Exposures | 62.5% | 35/56 |
| | b) Outcomes | 46.4% | 26/56 |
| | c) Confounders and effect modifiers | 16.1% | 9/56 |
| 12.1 | Authors should describe the extent to which the investigators had access to the database population used to create the study population. | 87.5% | 49/56 |
| 12.2 | Authors should provide information on the data-cleaning methods used in the study. | 1.8% | 1/56 |
| 12.3 | State whether the study included person level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided. | 64.3% | 36/56 |
| 13.1 | Describe in detail the selection of the persons included in the study (i.e., study population selection), including filtering based on data quality, data availability, and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram. | 85.7% | 48/56 |
| 19.1 | Discuss the implications of using data that were not created or collected to answer the specific research question(s). | | |
| | a) Include discussion of misclassification bias as they pertain to the study being reported. | 46.4% | 26/56 |
| | b) Include discussion of unmeasured confounding as they pertain to the study being reported. | 73.2% | 41/56 |
| | c) Include discussion of missing data as they pertain to the study being reported. | 82.1% | 46/56 |
| | d) Include discussion of changing eligibility over time as they pertain to the study being reported. | 14.3% | 8/56 |
| 22.1 | Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code. | 50.0% | 28/56 |

readily available electronic information will likely continue to drive RCD research. An a priori hypothesis and analytical plan, appropriate statistical techniques, a careful assessment of bias, and high-quality reporting will hopefully continue to improve the quality and impact of RCD studies. Despite the limitations of observation studies, they often produce results similar to randomized, controlled trials,[27] and observational studies using RCD can have a significant impact on healthcare, as evidenced by their ability to influence regulatory bodies.[28] Future changes in RCD reporting may include journals mandating the use of the RECORD checklist, the use of online registries for study protocols[29] (similar to randomized trial registration), or the publication of the actual data-coding and raw analytic results to improve transparency.

There are several exciting developments that are applicable to RCD. First is the advent of the registry-based randomized, controlled trial.[30] This combines the advantages of prospective, randomized trials with RCD. Patients are prospectively recruited, screened, and randomized at baseline, and important clinical characteristics are determined using traditional tools, such as history, physical exam, and review of medical records. Patient data is then deterministically linked with RCD to assess for variables not easily

captured in prospective studies (such as healthcare use or prior medication exposure). The patients are then "followed" for the trial outcomes using the existing RCD processes. This significantly reduces loss to followup, increases the appeal of the study to participants (as there are no study-related followup visits), and markedly reduces the study costs, as research personnel do not have to keep track of participants beyond their initial visit.

Second is the integration of true "big data" into the traditional RCD sources used for medical research. This includes data from internet search engines, credit card companies, environmental data, fitness trackers, and social media behaviour. Businesses already use this information to identify and market their products, and these profiles could further augment a clinical researcher's ability to understand lifestyle and environmental factors and their relationship with diseases.

Third is the use of networks of RCD. These networks consist of a collaborative group of investigators with the ability to access unique RCD sources. A common study protocol with strict analytic controls is created and applied across multiple RCD sources (including multiple regions and countries).[31] Pooled results with a consistent effect across different RCD data sets and countries can create a very impactful observational study with robust conclusions.[32]

Finally, RCD is available in many data sources as far back as the 1970s. This means that the first generation of people born and raised in an electronic healthcare environment are starting families of their own. Complex analytic questions around the inheritance of disease (using RCD linked across multiple generations), twin studies (identified using RCD rather than traditional twin cohorts), and the risk for future diseases based on a lifelong catalogue of medical encounters will be possible.

**Competing interests:** Dr. Welk has received an investigator-initiated grant from Astellas. Dr. Kwong reports no competing personal or financial interests.

This paper has been peer-reviewed.

## References

1. Schlomer BJ, Copp HL. Secondary data analysis of large data sets in urology: Successes and errors to avoid. *J Urol* 2014;191:587-96. https://doi.org/10.1016/j.juro.2013.09.091
2. PSA values and SEER data. http://seer.cancer.gov/data/psa-values.html. Published May 17, 2016. Accessed March 16, 2016.
3. Copeland KT, Checkoway H, McMichael AJ, et al. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol* 1977;105:488-95. https://doi.org/10.1093/oxfordjournals.aje.a112408
4. Höfler M. The effect of misclassification on the estimation of association: A review. *Int J Methods Psychiatr Res* 2005;14:92-101. https://doi.org/10.1002/mpr.20
5. Funk MJ, Landi SN. Misclassification in administrative claims data: Quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep* 2014;1:175-85. https://doi.org/10.1007/s40471-014-0027-z
6. Normand SLT, Sykora K, Li P, et al. Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ* 2005;330:1021. https://doi.org/10.1136/bmj.330.7498.1021
7. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. *Stat Med* 2014;33:1242-58. https://doi.org/10.1002/sim.5984
8. Guertin JR, Rahme E, Dormuth CR, et al. Head-to-head comparison of the propensity score and the high-dimensional propensity score-matching methods. *BMC Med Res Methodol* 2016;16:22. https://doi.org/10.1186/s12874-016-0119-1
9. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014;33:2297-340. https://doi.org/10.1002/sim.6128
10. Ørsted DD, Bojesen SE, Nielsen SF, et al. Association of clinical benign prostate hyperplasia with prostate cancer incidence and mortality revisited: A nationwide cohort study of 3 009 258 men. *Eur Urol* 2011;60:691-8. https://doi.org/10.1016/j.eururo.2011.06.016
11. Benchimol EI, Manuel DG, To T, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2010:1-9. https://doi.org/10.1016/j.jclinepi.2010.10.006
12. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. *J Clin Epidemiol* 2011;64:1054-9. https://doi.org/10.1016/j.jclinepi.2011.01.001
13. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol* 2012;65:126-31. https://doi.org/10.1016/j.jclinepi.2011.08.002
14. Begg C. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:637-9. https://doi.org/10.1001/jama.276.8.637
15. Elm von E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Epidemiology* 2007;18:800-4. https://doi.org/10.1097/EDE.0b013e3181577654
16. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885. https://doi.org/10.1371/journal.pmed.1001885
17. Nicholls SG, Quach P, Elm von E, et al. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for arriving at consensus and developing reporting guidelines. *PLoS ONE* 2015;10:e0125620. https://doi.org/10.1371/journal.pone.0125620.s002
18. da Costa BR, Cevallos M, Altman DG, et al. Uses and misuses of the STROBE statement: Bibliographic study. *BMJ Open* 2011;1:e000048. https://doi.org/10.1136/bmjopen-2010-000048
19. Hassett MJ, Ritzwoller DP, Taback N, et al. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using two large contemporary cohorts. *Med Care* 2014;52:e65-e73. https://doi.org/10.1097/MLR.0b013e318277eb6f
20. Semins MJ, Trock BJ, Matlaga BR. Validity of administrative coding in identifying patients with upper urinary tract calculi. *J Urol* 2010;184:190-2. https://doi.org/10.1016/j.juro.2010.03.011
21. Sewell JM, Rao A, Elliott SP. Validating a claims-based method for assessing severe rectal and urinary adverse effects of radiotherapy. *Urology* 2013;82:335-40. https://doi.org/10.1016/j.urology.2013.02.071
22. Desai RJ, Solomon DH, Shadick N, et al. Identification of smoking using Medicare data—a validation study of claims-based algorithms. *Pharmacoepidemiol Drug Saf* 2016;25:472-5. https://doi.org/10.1002/pds.3953
23. Muntner P, Gutiérrez OM, Zhao H, et al. Validation study of Medicare claims to identify older U.S. adults with CKD using the Reasons for Geographic and Racial Differences in Stroke (REGARDS) Study. *Am J Kid Dis* 2015;65:249-58. https://doi.org/10.1053/j.ajkd.2014.07.012
24. Springate DA, Kontopantelis E, Ashcroft DM, et al. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS ONE 2014;9(6):e99825-e99825. doi:10.1371/journal.pone.0099825. https://doi.org/10.1371/journal.pone.0099825
25. Van den Broeck J, Cunningham SA, Eeckels R, et al. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Med* 2005;2:e267. https://doi.org/10.1371/journal.pmed.0020267
26. Data Universe Explosion & the Growth of Big Data, CSC. Available at: http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode. Published May 24, 2016. Accessed May 24, 2016.
27. Anglemyer A, Horvath HT, Bero L. Healthcare Outcomes Assessed with Observational Study Designs Compared with Those Assessed in Randomized Trials. (Bero L, ed.). Chichester, UK: John Wiley & Sons, Ltd; 1996:1-46. https://doi.org/10.1002/14651858.MR000034.pub2
28. Williams RJ, Tse T, Harlan WR, et al. Registration of observational studies: Is it time? *CMAJ* 2010;182:1638-42. https://doi.org/10.1503/cmaj.092252

29. Loder E, Groves T, Macauley D. Registration of observational studies. *BMJ* 2010;340:c950. https://doi.org/10.1136/bmj.b4184
30. Lauer MS, D'Agostino RB. The randomized registry trial — the next disruptive technology in clinical research? *N Engl J Med* 2013;369:1579-81. https://doi.org/10.1056/NEJMp1310102
31. Suissa S, Henry D, Caetano P, et al. CNODES: The Canadian Network for Observational Drug Effect Studies. *Open Med* 2012;6:e134-e140. https://doi.org/10.1038/clpt.1985.188
32. Filion KB, Azoulay L, Platt RW, et al. A multicentre, observational study of incretin-based drugs and heart failure. *N Engl J Med* 2016;374:1145-54. https://doi.org/10.1056/NEJMoa1506115


**Correspondence:** Dr. Blayne Welk, Department of Surgery and Epidemiology and Biostatistics, Western University, London ON; bkwelk@gmail.com

**To answer the three multiple-choice questions associated with this review article, go to:**

**https://www.qzzr.com/c/quiz/378689/claim-section-3-self-assessment-credits-for-reading-cuaj-04-17**

This program is an Accredited Self-Assessment Program (Section 3) as defined by the Maintenance of Certification Program of The Royal College of Physicians & Surgeons of Canada, and approved by the Canadian Urological Association. Remember to visit MAINPORT (www.mainport.org/mainport/) to record your learning and outcomes. You may claim a maximum of 1 hour of credit.