The Journal of Open Source Software

# coder: An R package for code-based item classification and categorization

**Erik Bülow**[1, 2]

**1** The Swedish Arthroplasty Register, Registercentrum Västra Götaland, Gothenburg, Sweden **2** Department of Orthopaedics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

**DOI:** 10.21105/joss.02916

**Software**

- Review
- Repository
- Archive

**Editor:** Kristen Thyng

**Reviewers:**

- @kthyng

**Submitted:** 16 December 2020
**Published:** 18 December 2020



## Summary

The `coder` package lets researchers classify and categorize coded item data using pre-specified classification schemes based on regular expressions. Default classification schemes are included for commonly used medical and clinical classifications. The package implementation aim for high performance and the package can be used with large data sets.

## Medical coding and classifications

Registry based research and the use of real world evidence (RWE) and data (RWD) have gained popularity over the last years (Sherman et al., 2016), both as an epidemiological research tool, and for monitoring post market safety and adverse events due to regulatory decisions. Data from administrative, clinical and medical registries are often coded based on standardized classifications for diagnostics, procedures/interventions, medications/medical devices and health status/functioning.

Codes and classifications are maintained and developed by several international bodies, such as The World Health Organization (WHO), SNOMED International, and the Nordic Medico-Statistical Committee (NOMESCO).

## Challanges

Common classifications such as the International Classification of Diseases (ICD) or the Anatomical Therapeutic Chemical Classification System (ATC) entails thousands of codes which are hard to use and interpret in applied research. This is often solved by an abstraction layer combining individual codes into broader categories, sometimes further simplified by a single index value based on a weighted sum of individual categories (Charlson et al., 1987; Elixhauser et al., 1998; Pratt et al., 2018; Quan et al., 2005; Sloan et al., 2003).

## Statement of Need

Large and long-standing national databases often contain millions of entries and span several Gigabytes (GB) in size. This leads to high computational burden and a time-consuming data managing process, a cumbersome but necessary prerequisite before any relevant analysis can be performed. There are several R-packages with a deliberate focus on comorbidity data coded

by ICD and summarized by the Charlson or Elixhauser comorbidity indices (icd, comorbidity (Gasparini, 2018) and medicalrisk). The `coder` package includes such capabilities as well, but takes a more general approach to deterministic item classification and categorization.

## The coder package

coder is an R package with a scope to combine items (i.e. patients) with generic code sets, and to classify and categorize such data based on generic classification schemes defined by regular expressions. It is easy to combine different classifications (such as multiple versions of ICD, ATC or NOMESCO codes), with different classification schemes (such as Charlson, Elixhauser, RxRisk V or for example local definitions of adverse events after total hip arthroplasty) and different weighted indices based on those classifications. The package includes default classification schemes for all those settings, as well as an infrastructure to implement and visualize custom classification schemes. Additional functions simplify identification of codes and events within limited time frames, such as comorbidity during one year before surgery or adverse events within 30 days after. `coder` can also be used in tandem with decoder, a package facilitating interpretation of individual codes.

`coder` has been optimized for speed and large data sets using reference semantics from data.table, matrix-based computations and code profiling. The prevalence of large datasets makes it difficult to use parallel computing however, since the limit of available random-access memory (RAM) often implies a more serious bottleneck, which limits the possibility to manifold data sets for multiple cores.

`coder` has been used in ongoing, as well as in previously published research (Berg et al., 2018; Erik Bülow et al., 2020, 2017; E. Bülow et al., 2019; P. Cnudde et al., 2017; P. H. J. Cnudde et al., 2018; Hansson et al., 2020; Jawad et al., 2019; Nemes et al., 2018; Wojtowicz et al., 2019).

## References

Berg, U., Bülow, E., Sundberg, M., & Rolfson, O. (2018). No increase in readmissions or adverse events after implementation of fast-track program in total hip and knee replacement at 8 Swedish hospitals: An observational before-and-after study of 14,148 total joint replacements 2011–2015. *Acta Orthopaedica*, *89*(5), 522–527. https://doi.org/10.1080/17453674.2018.1492507

Bülow, E., Cnudde, P., Rogmark, C., Rolfson, O., & Nemes, S. (2019). Low predictive power of comorbidity indices identified for mortality after acute arthroplasty surgery undertaken for femoral neck fracture. *The Bone & Joint Journal*, *101-B*(1), 104–112. https://doi.org/10.1302/0301-620X.101B1.BJJ-2018-0894.R1

Bülow, Erik, Nemes, S., & Rolfson, O. (2020). Are the first or the second hips of staged bilateral THAs more similar to unilateral procedures? A study from the Swedish hip arthroplasty register. *Clinical Orthopaedics and Related Research*, 1. https://doi.org/10.1097/CORR.0000000000001210

Bülow, Erik, Rolfson, O., Cnudde, P., Rogmark, C., Garellick, G., & Nemes, S. (2017). Comorbidity does not predict long-term mortality after total hip arthroplasty. *Acta Orthopaedica*, *88*(July), 1–6. https://doi.org/10.1080/17453674.2017.1341243

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, *40*(5), 373–383. https://doi.org/10.1016/0021-9681(87)90171-8

Cnudde, P. H. J., Nemes, S., Bülow, E., Timperley, A. J., Whitehouse, S. L., Kärrholm, J., & Rolfson, O. (2018). Risk of further surgery on the same or opposite side and mortality after primary total hip arthroplasty: A multi-state analysis of 133,654 patients from the Swedish Hip Arthroplasty Register. *Acta Orthopaedica*, *89*(x). https://doi.org/10.1080/17453674.2018.1475179

Cnudde, P., Nemes, S., Bülow, E., Timperley, J., Malchau, H., Kärrholm, J., Garellick, G., & Rolfson, O. (2017). Trends in hip replacements between 1999 and 2012 in Sweden. *Journal of Orthopaedic Research*, *36*(January), 432–442. https://doi.org/10.1002/jor.23711

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, *36*(1), 8–27.

Gasparini, A. (2018). *Comorbidity: An R package for computing comorbidity scores software review repository archive*. https://doi.org/10.21105/joss.00648

Hansson, S., Bülow, E., Garland, A., Kärrholm, J., & Rogmark, C. (2020). More hip complications after total hip arthroplasty than after hemiarthroplasty as hip fracture treatment: Analysis of 5,815 matched pairs in the Swedish Hip Arthroplasty Register. *Acta Orthopaedica*, *91*(2), 133–138. https://doi.org/10.1080/17453674.2019.1690339

Jawad, Z., Nemes, S., Bülow, E., Rogmark, C., & Cnudde, P. (2019). Multi-state analysis of hemi- and total hip arthroplasty for hip fractures in the Swedish population - Results from a Swedish national database study of 38,912 patients. *Injury*, *50*(2), 272–277. https://doi.org/10.1016/J.INJURY.2018.12.022

Nemes, S., Lind, D., Cnudde, P., Bülow, E., Rolfson, O., & Rogmark, C. (2018). Relative survival following hemi-and total hip arthroplasty for hip fractures in Sweden. *BMC Musculoskelet Disord*, *19*(1), 407. https://doi.org/10.1186/s12891-018-2321-2

Pratt, N. L., Kerr, M., Barratt, J. D., Kemp-Casey, A., Ellett, L. M. K., Ramsay, E., & Roughead, E. E. (2018). The validity of the Rx-Risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system. *BMJ Open*, *8*(4). https://doi.org/10.1136/bmjopen-2017-021122

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., A Beck, C., Feasby, T. E., & A Ghali, W. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, *43*(11), 1130–1139. https://doi.org/10.1097/01.mlr.0000182534.19832.83

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., Shuren, J., Temple, R., Woodcock, J., Yue, L. Q., & Califf, R. M. (2016). Real-World Evidence What Is It and What Can It Tell Us? *N Engl J Med*, *375*(23), 2293–2297. https://doi.org/10.1056/NEJMsb1609216

Sloan, K. L., Sales, A. E., Liu, C.-F., Fishman, P., Nichol, P., Suzuki, N. T., & Sharp, N. D. (2003). Construction and characteristics of the RxRisk-V: A VA-adapted pharmacy-based case-mix instrument. *Medical Care*, *41*(6), 761–774. https://doi.org/10.1097/01.MLR.0000064641.84967.B7

Wojtowicz, A. L., Mohaddes, M., Odin, D., Bülow, E., Nemes, S., & Cnudde, P. (2019). Is Parkinson's Disease Associated with Increased Mortality, Poorer Outcomes Scores, and Revision Risk After THA? Findings from the Swedish Hip Arthroplasty Register: *Clinical Orthopaedics and Related Research*, *477*(6), 1347–1355. https://doi.org/10.1097/CORR.0000000000000679