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

Erik Bülow¹, 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

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.

Challenges

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

Bülow, E., (2020). coder: An R package for code-based item classification and categorization. Journal of Open Source Software, 5(56), 2916. https://doi.org/10.21105/joss.02916
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

Bülow, E. (2020). coder: An R package for code-based item classification and categorization. *Journal of Open Source Software, 5*(56), 2916. https://doi.org/10.21105/joss.02916
