International Journal of Population Data Science

Code List Library: A Solution to Improve Research Repeatability, Transparency, and Efficiency by Curating Lists of Clinical Codes

Thayer, D[1], Bown, D[1], Leake, T[1], Jones, J-L[1], Noyce, R[1], Brooks, C[2], and Ford, D[1]

[1]Swansea University

[2]SAIL Databank, Swansea University

Introduction

Sets of clinical codes that define conditions and events of interest are a key knowledge product in health data research. Documenting such lists is essential for transparency and repeatability, and there is great potential benefit in their sharing and reuse. We designed and implemented software to address these goals.

Objectives and Approach

Our goals were threefold:

1. Provide a graphical user interface (GUI) to allow easier creation of code lists, for less technical users.

2. Allow clear documentation of code lists, preserving the history of their creation and capturing metadata about their meaning, provenance, and use.

3. Facilitate programmatic access, so that the software is not just documentation but can be integrated into data preparation and analysis.

To these ends, we developed a web application using Python and PostgreSQL that allows creating, editing, and accessing via a GUI, as well as a REST API for integration into SQL, R, and other environments.

Results

The software allows users to view and create lists through a familiar web paradigm. Lists can be built by identifying codes in a variety of ways, including keyword searches, regular expressions, and more complex rules. A change history is stored.

Information such as a description, whether there was clinical reviewed, and relevant publications is captured.

The REST API allows access and use in a variety of settings. We have implemented a DB2 SQL interface to enable code lists to be used within database queries, and other interfaces such as an R package are planned for the future.

It will be used within the SAIL Databank initially, with a public version for sharing across institutions planned. The code will be open source to enable further development.

Conclusion/Implications

We expect this tool to facilitate faster, higher-quality, more reproducible research in Wales and beyond. Hopefully it will be not just a standalone effort, but one small piece in a set of better tools and methods that will enable our field to truly realized the benefit of large linked datasets.