International Journal of Population Data Science

Journal Website: www.ijpds.org

To Err is Human, all hail your Robot Overseer: automated testing as part of a robust data delivery platform.

Avery, D¹

¹University of Oxford

Introduction

We have a large cohort study of half a million people, which continually incorporates new data through health insurance, centre for disease control records, death certificates, resurveys, and ongoing quality assurance and participant information updates. To support our researchers we need data which is correct, up-to-date, and unchanging.

does introduce opportunities for both human and computer errors. Mitigating this risk with automated testing is convenient and effective.

Objectives and Approach

We must provide the new data, fixes and corrections to researchers, without missing anything or introducing issues. We make sequential iterations of our data available to researchers on a biannual basis; allowing a static version that can be referenced with regards to earlier work and providing the newest version of the data for new work. Due to the very large size of the data/code base and the small size of the team managing it, delivering this without error is a struggle. To mitigate this we developed testing scripts which catch issues and flag for resolution prior to release to researchers.

Results

We currently have 32 tests which catch all known issues which occur during a rebuild. On any occasion where a new type of issue is encountered, tests which would catch that issue and related issues are developed. As a result our last few releases have gone far more smoothly, with few if any issues reported after a release and certainly no previously encountered issues! Examples of current tests include: detection of a failed health insurance import; that we have the same number of participants; failure to increment version number between releases; checking that disease numbers have not changed dramatically over the shared timeframe.

Conclusion/Implications

Producing multiple static releases is a good way to balance the needs of a researcher for both static and current data, but it