Mortality Tracker - the COVID-19 case for real time Web APIs as Epidemiology Commons

Jonas S Almeida1,*, Meredith Shiels1, Praphulla Bhawsar1, Bhaumik Patel1, Erika Nemeth2, Richard Moffitt2, Montserrat Garcia Closas1, Neal Freedman1, Amy Berrington1

1Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA,
2Department of Biomedical Informatics, Stony Brook University (SUNY), Stony Brook NY, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: Mortality Tracker is an in-browser application for data wrangling, analysis, dissemination and visualization of public time series of mortality in the United States. It was developed in response to requests by epidemiologists for portable real time assessment of the effect of COVID-19 on other causes of death and all-cause mortality. This is performed by comparing 2020 real time values with observations from the same week in the previous 5 years, and by enabling the extraction of temporal snapshots of mortality series that facilitate modeling the interdependence between its causes.

Results: Our solution employs a scalable “Data Commons at Web Scale” approach that abstracts all stages of the data cycle as in-browser components. Specifically, the data wrangling computation, not just the orchestration of data retrieval, takes place in the browser, without any requirement to download or install software. This approach, where operations that would normally be computed server-side are mapped to in-browser SDKs, is sometimes loosely described as Web APIs, a designation adopted here.

Availability: https://episphere.github.io/mortalitytracker; webcast demo: youtu.be/ZsvCe7cZzLo.

Contact: jonas.dealmeida@nih.gov

1 Introduction

The COVID-19 pandemic is drastically changing interoperability expectations for epidemiological data systems. Both the scientific community and general public have real time expectations, which can lead to the challenge of loose adherence to data standards, fluid serialization norms, and uncertain QC at the source. The nature of such provisional data sources leads to constant updates by mechanisms that reflect the distribution and redundancy of the primary sources. Accordingly, data wrangling must be part of the real time continuous application delivery process, making time-sensitive data available, at scale, to a diverse community of experts across multiple domains.

By traversing over 20 million discharge records of New York state patients in real time, (Almeida et al., 2019) demonstrated the feasibility of a data wrangling architecture that leverages stateless APIs to fetch data from public Web Services. Their approach makes makes full use of advanced vectorized operations (map-reduce) in the web browser, improving the scalability of the “Data Commons at Web Scale” model. Tracking the net effect of COVID-19 on mortality overall and by cause (Weinberger et al., 2020) faced comparable challenges, further compounded by the real time requirement. The most salient of the new challenges is the provisional nature of data wrangling, namely, coping with data sources that change the classification of mortality count and evolve naming conventions and data structures, all without warning. Continuous deployment directly from versioning environment (Blischak et al., 2016) is therefore an essential aspect of the continuous delivery mechanism employed here. This is in contrast with conventional infrastructures, where the normalization is typically handled as a server-side component, as in (Wissel et al., 2020). A less obvious but critical advantage of full in-browser solutions is that they do not depend on server-side computation components: the execution environment and data access permissions are those of the researcher. This solution sidesteps intermediation with significant logistic requirements of persistence, liability, and compliance that most institutions hesitate to take in the midst of shifting research agendas and public health emergencies.
2 Methods

Mortality Tracker is a Web Application that provides a scalable client-side solution (entirely in JavaScript) that obviates the need to maintain less FAIR (Wilkinson et al., 2016) server-side components altogether. It relies on two real-time data sources, each of them a reference resource for tracking mortality in the United States: from different causes at the Centers for Disease Control (CDC), and from COVID-19, at both CDC and Johns Hopkins University (Dong et al., 2020).

At present, CDC classifies death certificates at weekly intervals in 13 causes of death ranging from cancer to COVID-19, and including both major chronic disease such as heart and kidney, and major communicable disease like influenza. These account for over 2/3 of all deaths, including approximately 1% assigned to a 14th “unclassified” cause. The detailed list of causes of death is provided with updated ICD-10 codes in the application described in this report, which includes links to supplementary material, data sources, and downloadable data extracts for each and every user-generated plot. Finally, alongside the link to the open source code, a community-facing chat service for questions and answers is embedded.

The Web application compares weekly deaths in 2020 with the previous five years in order to estimate the raw number of excess deaths per week, as well as cumulatively (Fig A,C). The user can select from the causes of death described for a specific jurisdiction (Fig C,D), or the entire USA (Fig A,B). For each selection of cause, jurisdiction and units, a context layered plot with all causes is also generated (Fig B). The data is retrieved in real time from CDC’s Socrata platform web services, SODA, by a client-side SDK developed along the principles and purpose described in (Almeida et al., 2019). This Web API also wrangles Johns Hopkins data from a single file (Dong et al., 2020). We plan to continue to add features to the mortality tracker as CDC releases more granular data, for example, on age, race/ethnicity and place of death for these different causes.

3 References

Almeida,J.S. et al. (2019) Serverless OpenHealth at data commons scale-traversing the 20 million patient records of New York’s SPARCS dataset in real time. PeerJ, 7, e6230.

Blischak,J.D. et al. (2016) A Quick Introduction to Version Control with Git and GitHub. PLoS Comput. Biol., 12, e1004668.

Dong,E. et al. (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis., 20, 533–534.

Weinberger,D.M. et al. (2020) Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020. JAMA Intern. Med.

Wilkinson,M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 3, 160018.

Wissel,B.D. et al. (2020) An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in real time. Journal of the American Medical Informatics Association.

Fig. 1. Snapshots of graphic displays produced by Mortality Tracker. Examples of A: Data aggregation and cumulative additional mortality across all jurisdictions with complete records, using averages from previous 5 years as baseline; B: the same representation for each primary cause reported by CDC layered, with Johns Hopkins Univ COVID-19 data superimposed at the bottom; C: Comparing real time COVID-19 mortality time series from the two sources for a single state (Maryland); D: example of a representation for a single cause of death (heart disease) and a single jurisdiction (NY City). At the foot of each interactive plot, links are provided to download the corresponding data in JSON and CSV format (not shown here).