daiquiri: Data Quality Reporting for Temporal Datasets

T. Phuong Quan [1¶]

1 University of Oxford, UK ¶ Corresponding author

DOI: 10.21105/joss.05034

Software

- Review
- Repository
- Archive

Editor: Arfon Smith

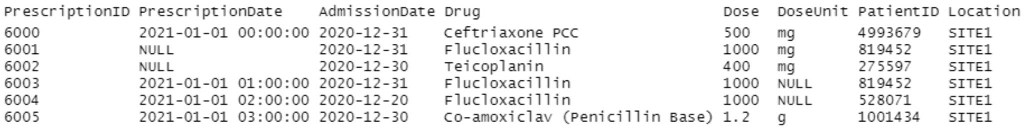
Reviewers:

- @arfon

Submitted: 13 December 2022
Published: 26 December 2022

Summary

The `daiquiri` R package generates data quality reports that enable quick visual review of temporal shifts in record-level data. It is designed with electronic health records in mind, but can be used for any type of record-level temporal data (i.e. tabular data where each row represents a single "event", one column contains the "event date", and other columns contain any associated values for the event, see Figure 1 for an example).



```
PrescriptionID PrescriptionDate    AdmissionDate Drug                         Dose  DoseUnit PatientID Location
6000           2021-01-01 00:00:00 2020-12-31    Ceftriaxone PCC              500   mg       4993679   SITE1
6001           NULL                2020-12-31    Flucloxacillin               1000  mg       819452    SITE1
6002           NULL                2020-12-30    Teicoplanin                  400   mg       275597    SITE1
6003           2021-01-01 01:00:00 2020-12-31    Flucloxacillin               1000  NULL     819452    SITE1
6004           2021-01-01 02:00:00 2020-12-20    Flucloxacillin               1000  NULL     528071    SITE1
6005           2021-01-01 03:00:00 2020-12-30    Co-amoxiclav (Penicillin Base) 1.2 g        1001434   SITE1
```

Figure 1: Example dataset containing information on antibiotic prescriptions.

The package automatically creates time series plots showing aggregated values for each data field (column) depending on its contents (e.g. min/max/mean values for numeric data, no. of distinct values for categorical data), see Figure 2, as well as overviews for missing values, non-conformant values, and duplicated rows, see Figure 3.

The resulting html reports are shareable and can contribute to forming a transparent record of the entire analysis process.

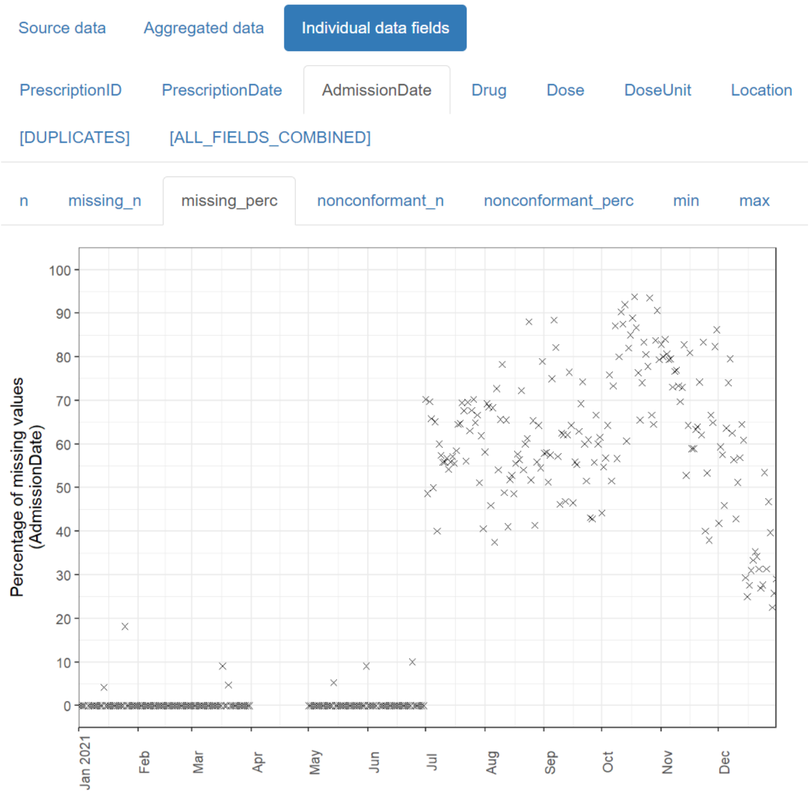


Figure 2: Screenshot showing percentage of missing values per day, for the AdmissionDate field of the example dataset.

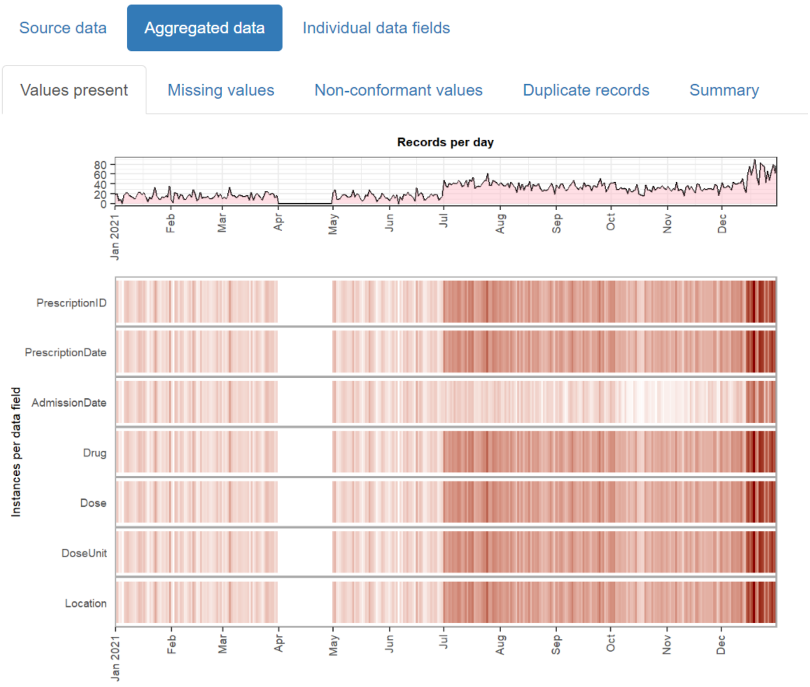


Figure 3: Screenshot showing number of values present per day, across all fields of the example dataset.

Statement of need

Large routinely-collected datasets are increasingly being used in research. However, given their data are collected for operational rather than research purposes, there is a greater-than-usual need for them to be checked for data quality issues before any analyses are conducted. Events occurring at the institutional level such as software updates, new machinery or processes can cause temporal artefacts that, if not identified and taken into account, can lead to biased results and incorrect conclusions.

For example, Figure 4 shows the mean value of all laboratory tests checking for levels of creatinine in the blood, from a large hospital group in the UK. As you can see, there are points in time where these values shift up or down suddenly and unnaturally, indicating that something changed in the way the data was collected or processed. A careful researcher needs to take these sudden changes into account, particularly if comparing or combining the data before and after these 'change points'.

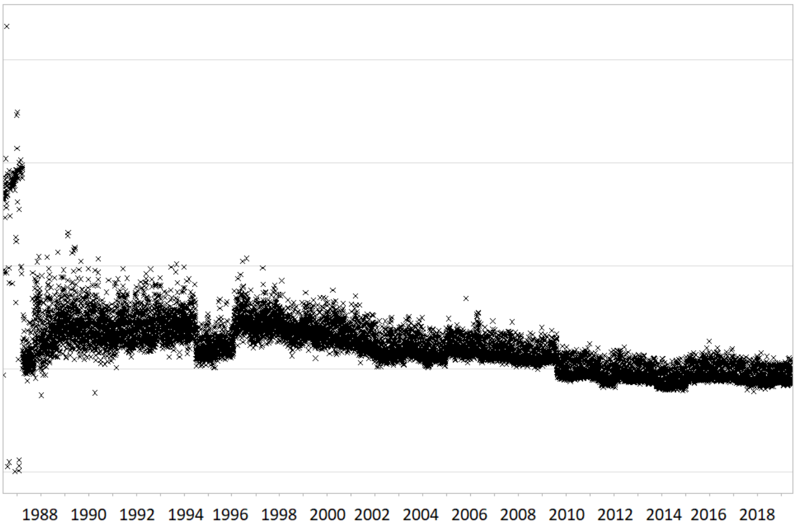


Figure 4: The mean value per day, of all laboratory tests checking for levels of creatinine in the blood.

While these checks should theoretically be conducted by the researcher at the initial data analysis stage, in practice it is unclear to what extent this is actually done, since it is rarely, if ever, reported in published papers. With the increasing drive towards greater transparency and reproducibility within the scientific community, this essential yet often-overlooked part of the analysis process will inevitably begin to come under greater scrutiny. The `daiquiri` package helps researchers conduct this part of the process more thoroughly, consistently and transparently, hence increasing the quality of their studies as well as trust in the scientific process.

There are a number of existing R packages which generate reports that provide an overview of a dataset's contents, such as `dataReporter` (formerly `dataMaid`) (Petersen & Ekstrøm, 2019), `smartEDA` (Putatunda et al., 2019), and `dataquieR` (Richter et al., 2021). In these packages, summary statistics are calculated across all rows in the dataset, or perhaps stratified by a categorical field. In contrast, `daiquiri` focuses on how these summary statistics may change over the time scale of the dataset, which can reveal data quality issues that might otherwise be missed when using these other packages.

Acknowledgements

This package was created as part of a PhD project, supervised by A. Sarah Walker, Tim Peto, Martin Landray, and Ben Lacey.

I would like to thank the rOpenSci team for their valuable input.

This work was supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) (NIHR200915), and by the NIHR Oxford Biomedical Research Centre.

References

Petersen, A. H., & Ekstrøm, C. T. (2019). dataMaid: Your assistant for documenting supervised data quality screening in r. *Journal of Statistical Software*, *90*(6), 1–38. https://doi.org/10.18637/jss.v090.i06

Putatunda, S., Ubrangala, D., Rama, K., & Kondapalli, R. (2019). SmartEDA: An r package for automated exploratory data analysis. *Journal of Open Source Software*, *4*(41), 1509. https://doi.org/10.21105/joss.01509

Richter, A., Schmidt, C. O., Krüger, M., & Struckmann, S. (2021). dataquieR: Assessment of data quality in epidemiological research. *Journal of Open Source Software*, *6*(61), 3093. https://doi.org/10.21105/joss.03093