# Understanding ehrQL Feature Usage in the OpenSAFELY Platform for Health Research

By: Vivian Okafor

okaforvivian4@gmail.com

## BACKGROUND

ehrQL (rhymes with circle; the Electronic Health Record Query Language), simplifies working with health data and has been used in around 75 research projects on OpenSAFELY, an open-source platform that securely supports analysis of over 58 million patient records in England. This project examined how researchers use ehrQL by identifying the most and least frequently used features, checking for reliance on outdated functions, and assessing how often ehrQL code is actively tested. The findings provide insights to improve the platform, encourage best practice, and guide future training and documentation.

## HOW?

The project followed a light step-by-step process, similar to an ETL pipeline (Extract → Transform → Load):

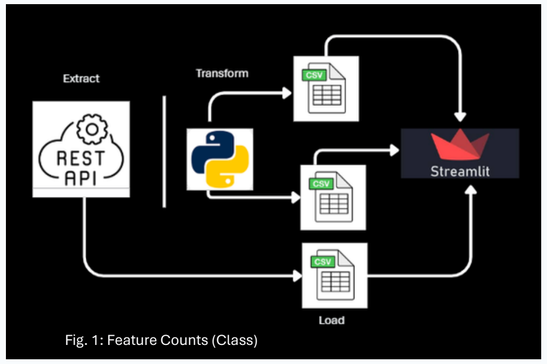


Fig. 1: Feature Counts (Class)

**Step 1 – Extract:** I used the GitHub REST API to collect all OpenSAFELY repositories and identify Python files containing ehrQL code.

**Step 2 – Transform:** Then parsed the files with Python, counted ehrQL feature usage, and mapped features to repositories.

**Step 3 – Load:** Saved outputs in two CSVs – one with total feature counts, another linking features to repositories.

**Visualisation:** Built an interactive Streamlit dashboard to show repositories, feature counts, and feature-by-repository breakdowns.

## RESULTS

Looking at how ehrQL is used across OpenSAFELY repositories reveals some clear trends in how researchers handle electronic health record data. Features like where() and codelist_from_csv() are used a lot, showing how important they are for filtering datasets and defining patient cohorts. Temporal and aggregation features are also common, reflecting the need for patient-level and longitudinal analyses.

On the other hand, some features are rarely or never used, perhaps because they are less relevant, not well known, or redundant.

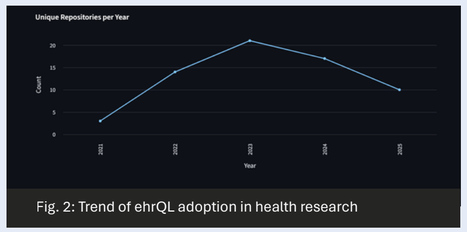


Fig. 2: Trend of ehrQL adoption in health research

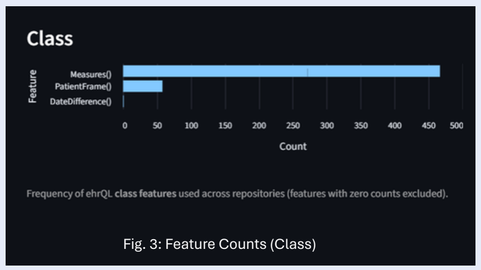


Fig. 3: Feature Counts (Class)

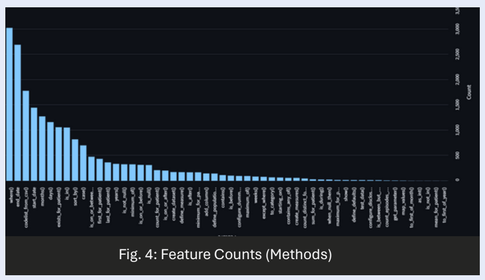


Fig. 4: Feature Counts (Methods)

## RELEVANCE

This project has several important implications for healthcare research. From the results, it's clear that focusing training and documentation on the features researchers use most can make their work more efficient, helping them spend less time troubleshooting or learning rarely used features.

Using core features consistently also improves reproducibility, so analyses across studies are more reliable. An understanding of which features are used the most can help the tech team prioritise platform improvements, simplify or retire underused features, and better support researchers' workflows. Altogether, these improvements can speed up research and generate actionable healthcare insights that inform policy, clinical practice, and patient care.

## WHAT'S NEXT?

This project has given a first look at how ehrQL is used in OpenSAFELY repositories, but there's plenty more that could be explored. Some of the next steps could include:

- Linking the analysis with job history to see:
  - When a repository first ran any ehrQL code.
  - Which features are most used in code that's been run, rather than just written
  - Which features appear in repositories but have never been run
  - How feature use has changed over time.

- Sorting repositories by when they first (or last) ran ehrQL code to spot trends in adoption and ongoing use.

I have written a blog post about my findings. Visit https://www.bennett.ox.ac.uk/blog/2025/08/understanding-ehrql-feature-usage-in-opensafely-for-health-research/

## ACKNOWLEDGEMENTS

I would like to thank the Bennett Institute and the HDR UK Health Data Science Internship programme for the opportunity to work on this project. Special thanks to Iain Dillingham, Peter Inglesby, Lucy Bridges, Brian Mackenna, Alice Wong, Providence Onyenekwe, and the entire Tech Team at the Bennett Institute for their guidance and support throughout my research.

## REFERENCE

OpenSAFELY. (n.d.). ehrQL documentation. Available at: https://docs.opensafely.org/ehrql/