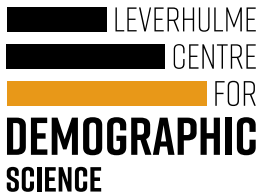


Adventures in Open and Assumption-Free Observational Research

Charlie Rahal

LCDS and Nuffield College, University of Oxford

A talk to AHRU, February 11th 2025



These slides are made with an open source typesetting language!

I_make_this_slide.tex

```

\begin{frame}[]
  These slides are made with an open source typesetting language!
  \begin{center}
    \begin{scriptsize}
      \begin{minipage}{.95\linewidth}
        \begin{myenv}{guess$_\_what$_\_this$_\_does.py}
          \begin{minted}{latex}
            # Hello friends :)
          \end{minted}
        \end{myenv}
      \end{minipage}
    \end{scriptsize}
  \end{center}
\end{frame}

```

General purpose programming in the form of Python:

guess_what_this_does.py

```
import json, io, imageio, requests
from PIL import Image

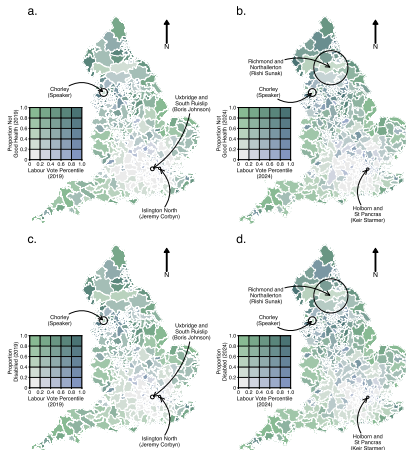
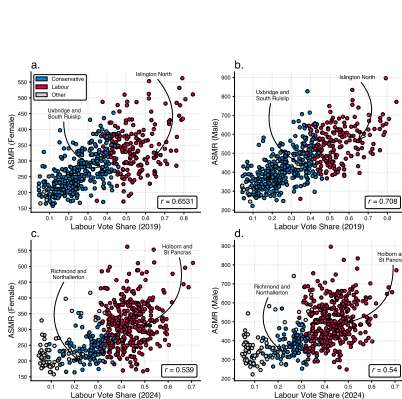
def get_cat_bounded(url):
    api_ret = requests.get(url)
    if api_ret.status_code==200:
        text = json.loads(api_ret.text)
        while (text[0]['width']<=400 or text[0]['width']>=600) or \
            (text[0]['height']<=400 or text[0]['height']>=600):
            api_ret = requests.get(url)
            if api_ret.status_code==200:
                text = json.loads(api_ret.text)
        r = requests.get(api_text[0]['url'])
        return Image.open(io.BytesIO(r.content))

url = 'https://api.thecatapi.com/v1/images/search'
images = []
for _ in range(0, 50):
    images.append(get_cat_bounded(url))
imageio.mimsave('CATS.gif', images, fps=1)
```

Thats right: Cats!



DataViz: Everywhere And All The Time.



‘Voting Patterns, Mortality, and Health Inequalities in England’

With Melinda C. Mills, Philip Clarke, Veline L'Esperance and Josh Knight

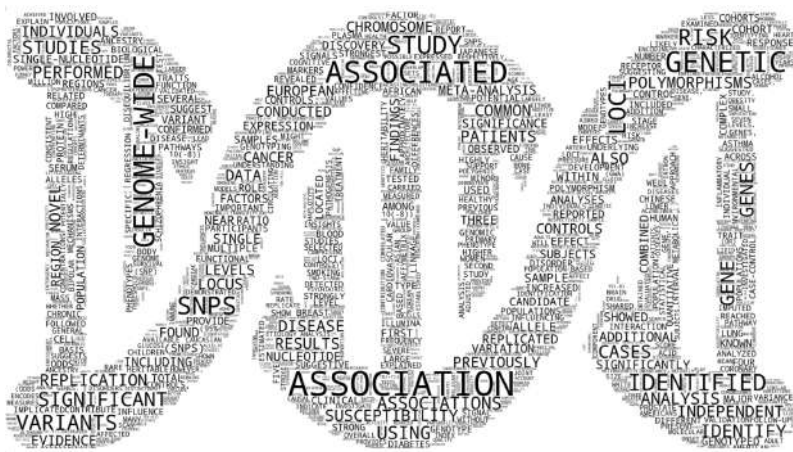
A lot of my work is ‘**scientometric**’/‘**bibliometric**’ in nature.

- **Scientometrics** is the field of study which concerns itself with measuring and analysing science.
- **Bibliometrics** is the use of statistical methods to analyze quantitative information pertaining to books, articles and other publications.

Through ‘Open Approaches’ to the scientific record, we can efficiently conduct scientometric reviews.

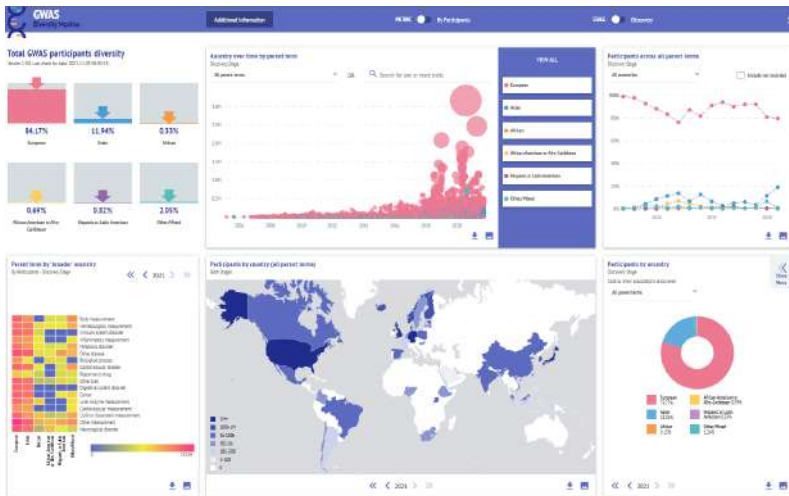
Studying Inequalities through Open Science

- My early scientometric work analyzed GWAS.
- It took curated lists of catalogued information from the European Bioinformatics Institute. We published two papers which:
 1. Conduct a thorough scientometric review.
 2. Launch a substantial interactive dashboard.
- There is **HUGE** scope for 'Care' research here.
 - Why is 'open' scientometrics good? It helps us understand what has already been done, and where inequalities remain!



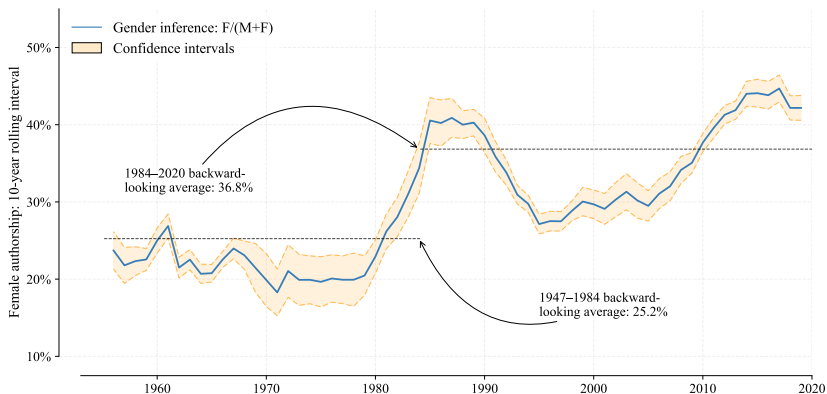
'The Scientometrics of Genome Wide Association Studies'

With Melinda C. Mills

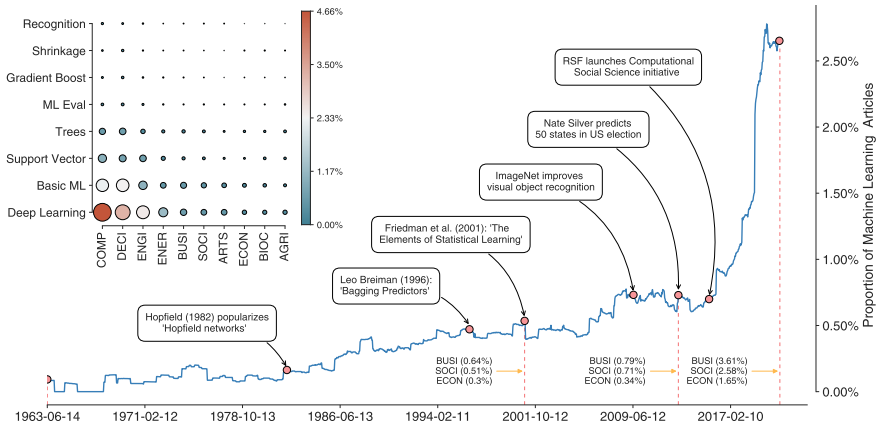


‘The GWAS Diversity Monitor Tracks Diversity in Real Time’
with Melinda C. Mills

Studying Gender with Open Science Tools

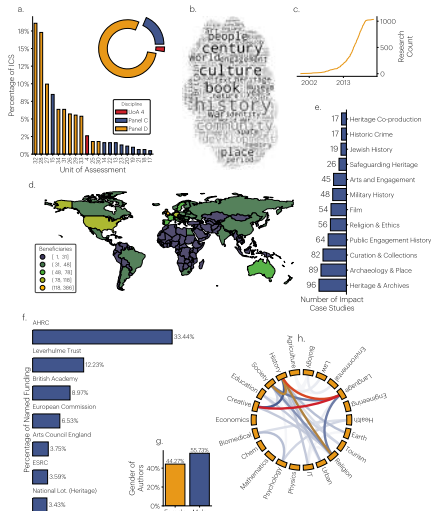
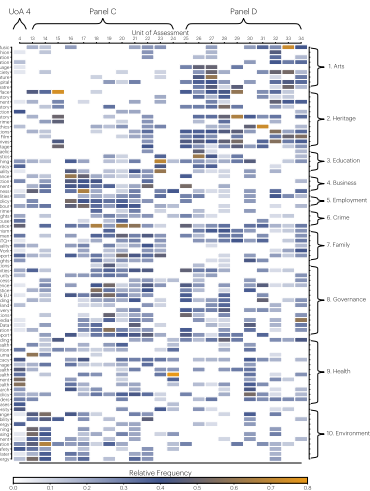


‘Population Studies at 75 Years: An empirical review’
With Melinda C. Mills



'The Rise of Machine Learning in the Academic Social Sciences'

With Dave Kirk and Mark Verhagen



'The SHAPE of Research'

With Sander Wagner, Melinda Mills, et al. (2024)

What's Next?

Two things coming up in this space:

1. Twenty Years of the UK Biobank;
2. ERC CoG to launch: 'The Evolution of Science'.



‘The Evolution of Science’ dashboard: An artists’s impression.

Administrative Data is an invaluable resource for Good.

ADR UK (2018)

- UK leads the world in terms of ‘Openness’ of government data.
- The value of public government data is increasingly recognised.
 - Increasingly recognised by the ESRC (e.g. ADRNs).
- For better or worse:
 - I’m more interested in ‘hard to use’, semi-structured data.

Administrative Data is an invaluable resource for Good.

ADR UK (2018)

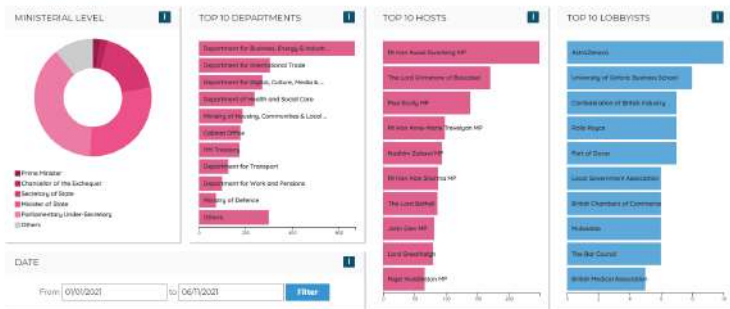
- UK leads the world in terms of ‘Openness’ of government data.
- The value of public government data is increasingly recognised.
 - Increasingly recognised by the ESRC (e.g. ADRNs).
- For better or worse:
 - I’m more interested in ‘hard to use’, semi-structured data.
 - Using curated data from the ONS SRS would be too easy.

TRANSPARENCY INTERNATIONAL UK Help | Contact Us Open Access Name Other versions **DONATE** About

2,623 out of 76,488 meetings **SEARCH BY LOBBYIST, HOST, SUBJECT** **Reset**  

UK GOVERNMENT

LOBBYIST



openaccess.transparency.org.uk

With Ian Knowles, Paulo Serôdio, David Stuckler and TI UK

Procurement Data

- David Cameron opened data on government contracts (2010).
- This provided a totally new type of semi-structured data.
- From from local authorities, to the NHS, to central government.
- This sounds like it might be useful for academic research...
- However, the raw data is **unbelievably** messy.

Procurement Data

- David Cameron opened data on government contracts (2010).
- This provided a totally new type of semi-structured data.
- From from local authorities, to the NHS, to central government.
- This sounds like it might be useful for academic research...
- However, the raw data is **unbelievably** messy.

If only there was a **Government Data Service** to curate it...

Procurement Data

- David Cameron opened data on government contracts (2010).
- This provided a totally new type of semi-structured data.
- From from local authorities, to the NHS, to central government.
- This sounds like it might be useful for academic research...
- However, the raw data is **unbelievably** messy.

If only there was a **Government Data Service** to curate it...

However, I am currently working with the ONS to improve collection!

Lets list a few of the problems with procurement data. Some are common to semi-structured data, some are not:

1. Some organisations don't believe in their legal obligation.
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations on the internet.)
3. Data commonly converted from a tabular file...

Lets list a few of the problems with procurement data. Some are common to semi-structured data, some are not:

1. Some organisations don't believe in their legal obligation.
2. Every file is in a heterogeneous format in a different location.
(There are **thousands** of locations on the internet.)
3. Data commonly converted from a tabular file... **into** a .pdf!
(This is a **direct** downgrade in the Berners-Lee 5* schema.)
4. Naturally, of course, there are no supplier UIDs.

Just one quick example, followed by what we can do...

A3Y3V

Expenditure Over Threshold

[illegible]

A3Y3V

Expenditure Over Threshold

[illegible]

A3Y3V

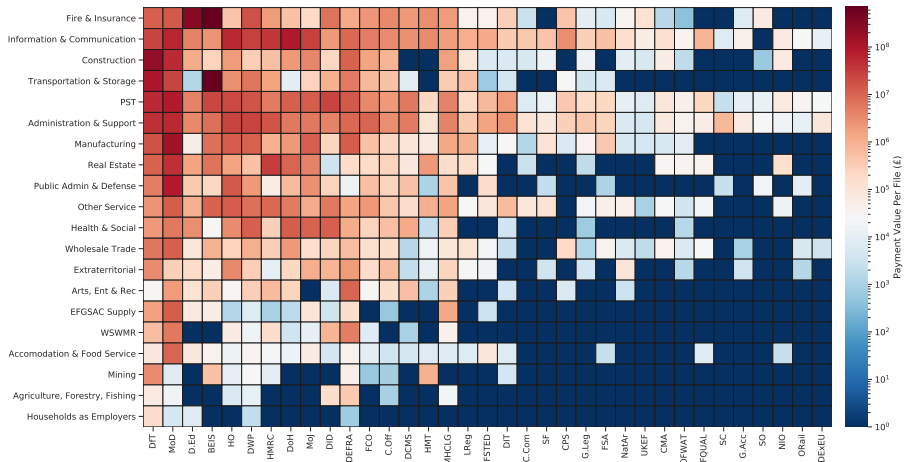
Expenditure Over Threshold

[illegible]

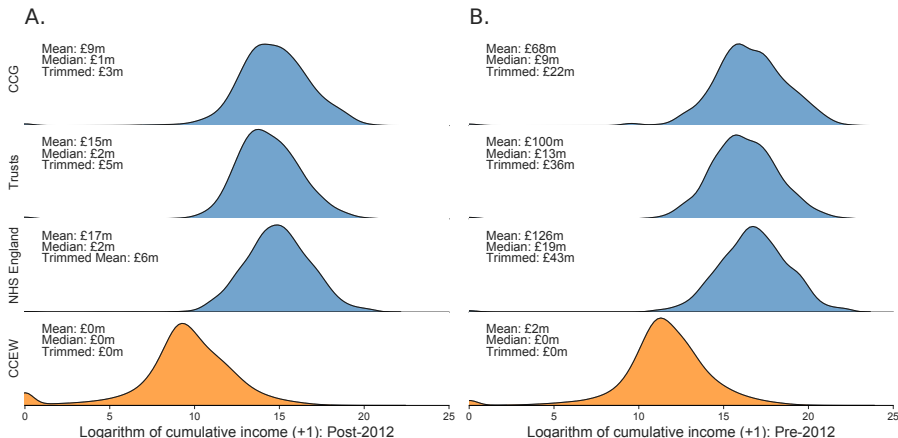
A3Y3Y

Expenditure Over Threshold

[illegible]



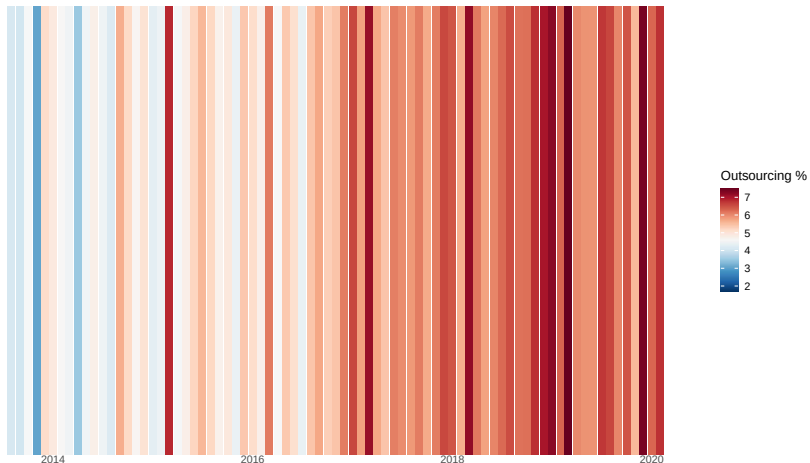
'Tools for Transparency in Central Government Spending'



‘The Role of Non-Profits in Public Health Service Provision’

With John Mohan

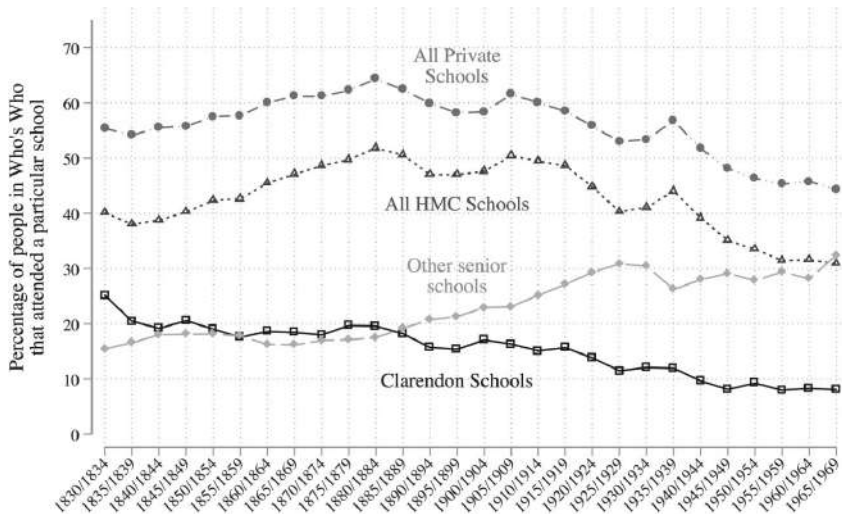
Creeping Privatisation in England's NHS



Source: Raw CCG payments approximately matched to Companies House and the Charity Commission, then manually verified

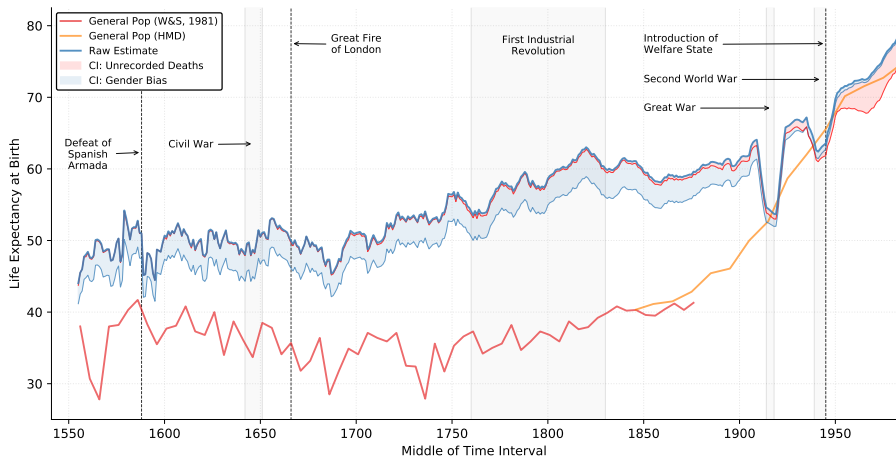
‘Creeping Commercialization in the NHS’

By Ben Goodair and Aaron Reeves



‘The decline and persistence of the old boy’

With Aaron Reeves, Sam Friedman. and Magne Flemmen



‘The Legacy of Longevity: Inequalities in UK Life Expectancy 1500-2016’ with Aaron Reeves, Felix Tropsch and Darryl Lundy

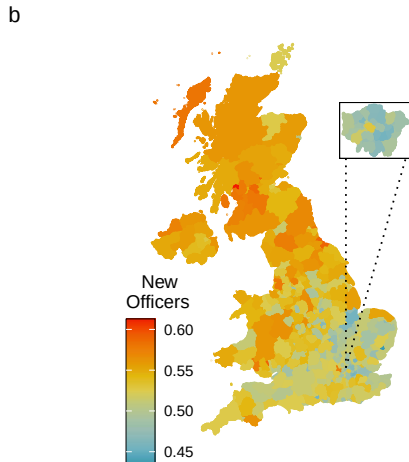
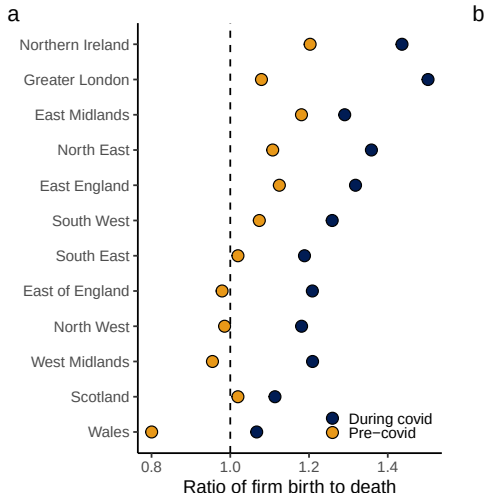
[illegible]

Ongoing research (With Aaron Reeves and Naomi Muggleton) uses and probate records from 1858-1996.

The idea is to:

1. Scrape 800k .pdfs containing 24m probates: **Done!**
2. Preprocess images ready for OCR: **Done!**
3. Run OCR on HPC in parallel: **Done!**
4. Build lifetables to calculate e_0 by wealth at death...: **Pending!**

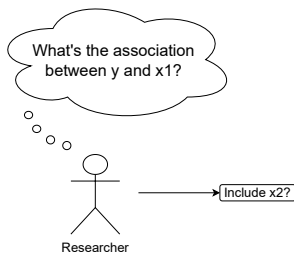
Capitalising on a Crisis

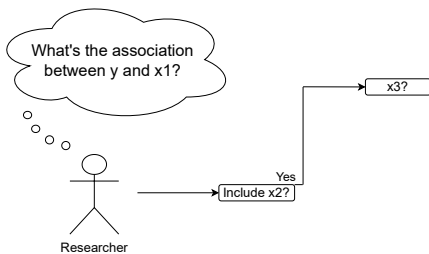


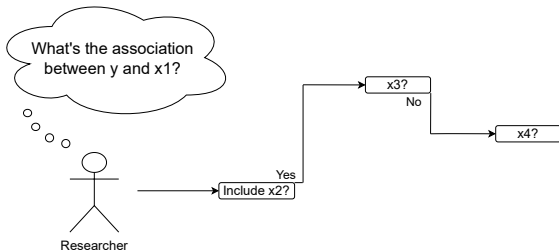
- There are close links between ‘Open Science’ and traditional modeling paradigms!
- By **opening** research, we can examine the certainty of statistical results.
- Does β_1 really equal 7.12452?

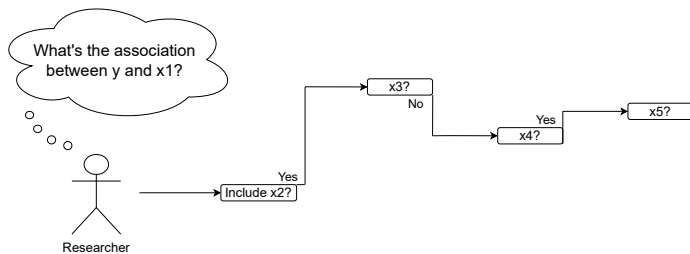
- There are close links between ‘Open Science’ and traditional modeling paradigms!
- By **opening** research, we can examine the certainty of statistical results.
- Does β_1 really equal 7.12452?
- How can we "take the ‘con’ out of econometrics"?

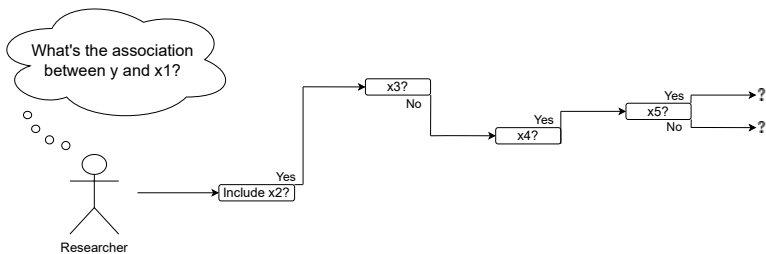
- There are close links between ‘Open Science’ and traditional modeling paradigms!
- By **opening** research, we can examine the certainty of statistical results.
- Does β_1 really equal 7.12452?
- How can we "take the ‘con’ out of econometrics"?
- Transparency and more computational procedures!

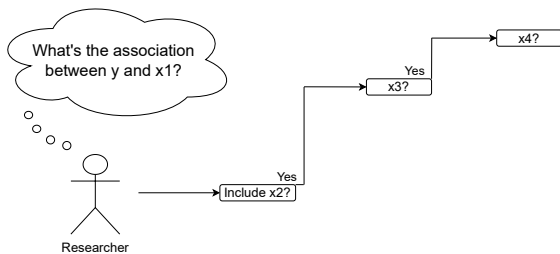


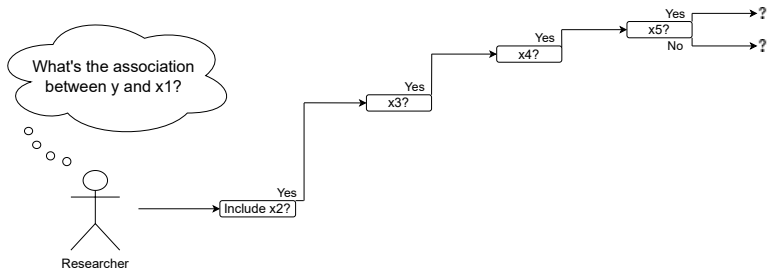


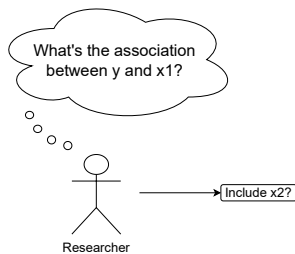


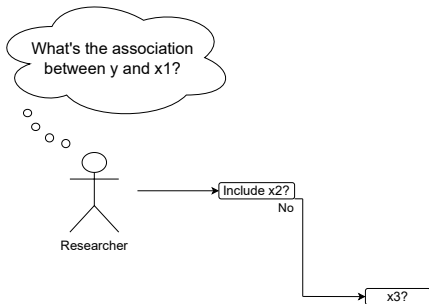


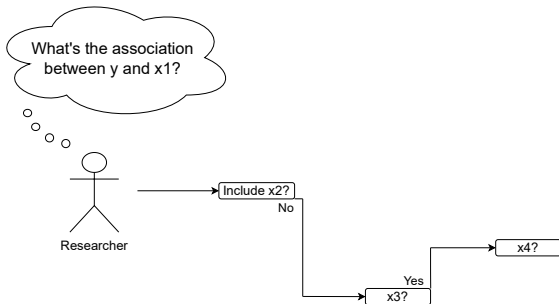


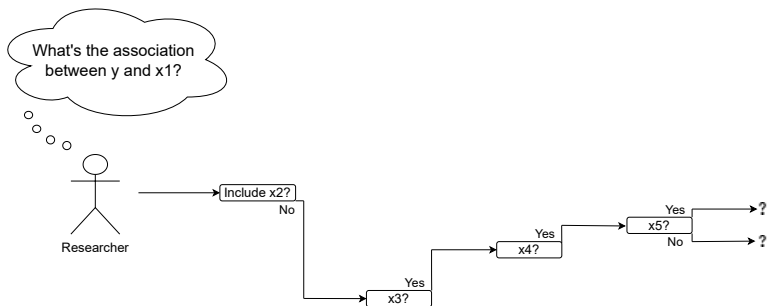


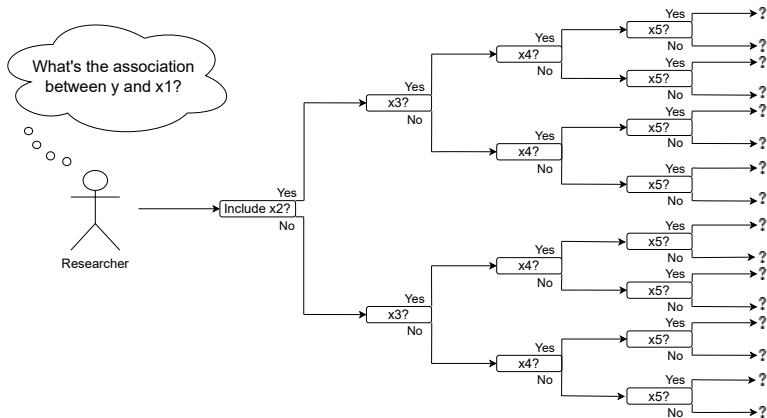




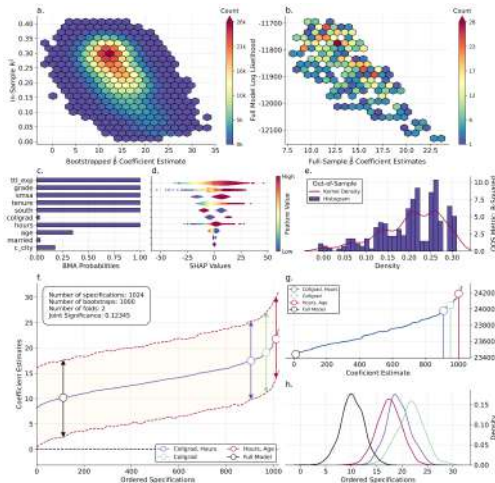








CfC research project with DVI and JY



1. Make modelling uncertainty more transparent!
2. Incorporate all multiverse tools in one place!
3. Allow researchers to do more rigorous quantitative research!



A simple, minimal website for RobustiPy!

[View the Project on GitHub here!](#)

[View the Python Documentation here!](#)

[View the RobustiPy working paper here!](#)

Welcome to the RobustiPy homepage!

RobustiPy is an efficient multiversal library with model selection, averaging, resampling and out-of-sample analysis. It analyses various output spaces, in addition to the control variable space (e.g. multiple dependent variables, estimands of interest, etc.). Developed for Python and R, it is designed to be both accessible and computationally efficient.

Note: This site (and project!) are a work in progress!

Installation

Installation is as simple (in Python) as:

```
pip install robustipy
```

and - potentially in the future for R - as:

```
install.packages("robustipy")
```

Examples: A series of toy examples with which might be useful to get you up and running can be found [here](#). A series of empirical examples which might be useful can be found [here](#).

Hackathon: We are holding a hackathon regarding RobustiPy in Oxford during June 2024. Participants can click [here](#) for further information on the day's activities!

Inspiration: We found great inspiration in the following three papers, which you might also enjoy reading:

Estimating the Cost of Informal Care: A novel two-stage approach to individual synthetic control

Maria Petrillo^{1,2*}, Daniel Valdenegro Ibarra^{1,3*}, Charles Rahal^{1,3,4}, Yanan Zhang^{1,5,6},
Gwilym Pryce^{1,7}, and Matthew R. Bennett^{1,8}

¹ESRC Centre for Care

²CIRCLE, University of Sheffield

³Leverhulme Centre for Demographic Science, University of Oxford

⁴Nuffield College, University of Oxford

⁵Oxford Institute of Population Aging, University of Oxford

⁶University College, University of Oxford

⁷Department of Economics, University of Sheffield

⁸School of Social Policy, University of Birmingham

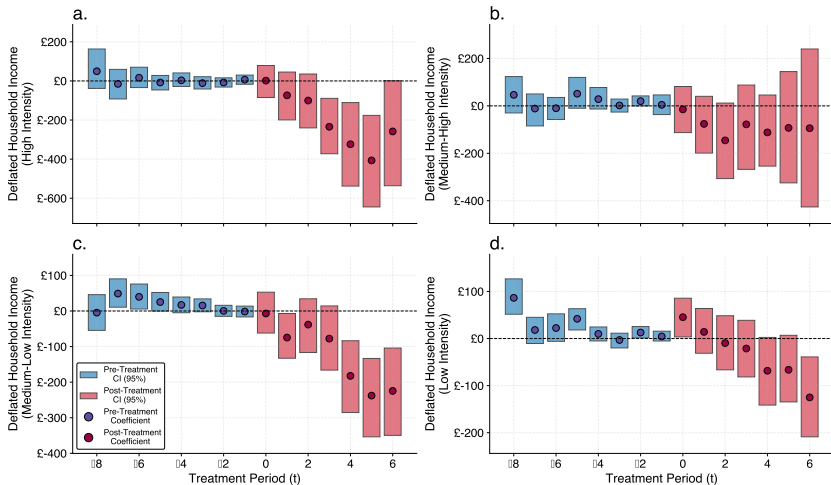
*Denotes joint lead authorship

September 19, 2024

Abstract

Informal carers provide the majority of care for people living with challenges related to older age, long-term illness, or disability. However, the care they provide carries a potentially significant income penalty for caregivers, a factor that is largely overlooked in both the economics literature and in policy discourse. Leveraging data from the UK Household Longitudinal Study, this paper provides the first robust causal estimates of the caregiving income penalty using a novel individual synthetic control method that accounts for unit-level heterogeneity in post-treatment trajectories over time. Our baseline estimates identify an average relative income gap of up to 45%, with an average decrease of £162 in monthly income, peaking at £192 per month after 4 years, based on the difference between informal carers providing the highest-intensity of care and their synthetic counterparts. We find that the income penalty is more pronounced for women than for men, and varies by ethnicity and age.

Keywords: *Causal Methods, Informal Care, Social Care, Longitudinal Analysis.*



Public

[Edit Pins](#)
[Unwatch](#)
[Fork](#)
[Star](#)

[main](#)
[2 Branches](#)
[0 Tags](#)

[Add file](#)
[Code](#)

MPetrillo25
Create psrm

e06c0b1 · 3 days ago
87 Commits

📁 .vscode	small analysis changes	6 months ago
📁 Stata_code	Create psrm	3 days ago
📁 figs	changes made up to paa 2024	4 months ago
📁 output/figures	changes made up to paa 2024	4 months ago
📁 outputs	changes made up to paa 2024	4 months ago
📁 replication_attempt	Reorganising to move stuff out of root	8 months ago
📁 src	changes made up to paa 2024	4 months ago
📄 .gitignore	Reorganising to move stuff out of root	8 months ago
📄 README.md	Update README.md	2 years ago

README

Cost of Care

This repository documents our work on calculating the cost of providing informal care. It takes an intersectional, comparative approach based on the [Individual Synthetic Control Method of Yagmi and Breen \(2021\)](#).

About

No description, website, or topics provided.

- 📖 Readme
- 🔗 Activity
- 📄 Custom properties
- ★ 2 stars
- 👁 2 watching
- 🍴 0 forks

[Report repository](#)

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Contributors

- dhvalden** Daniel Valdenegro Ibarra
- crahal**
- MPetrillo25**

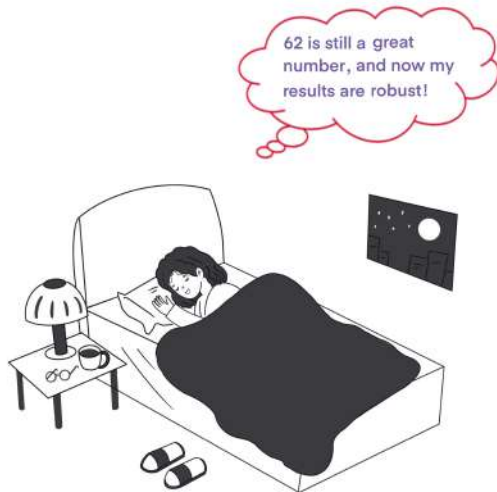


















Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur **everywhere!**

Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur **everywhere!**
- The variation in estimand can be **huge**.

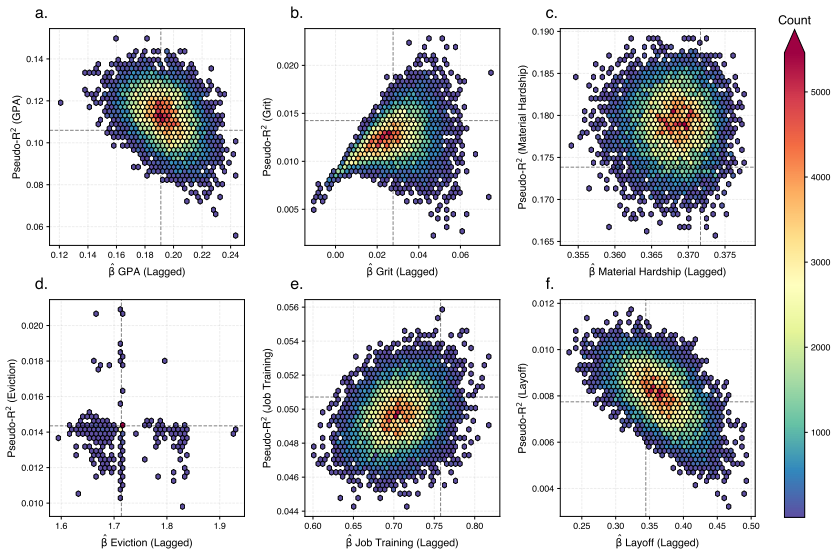
Pseudo-Random Number Generation (Cont.)

- We want to consider possible variation independent of arbitrary variation in seed choice.
- This is an extremely important and scarcely research problem.
- Pseudo-Random Number Generators occur **everywhere!**
- The variation in estimand can be **huge**.
- We bring attention to this through multiple types of replications.
 - Simulations, machine learning, topic modelling, and inferential research.

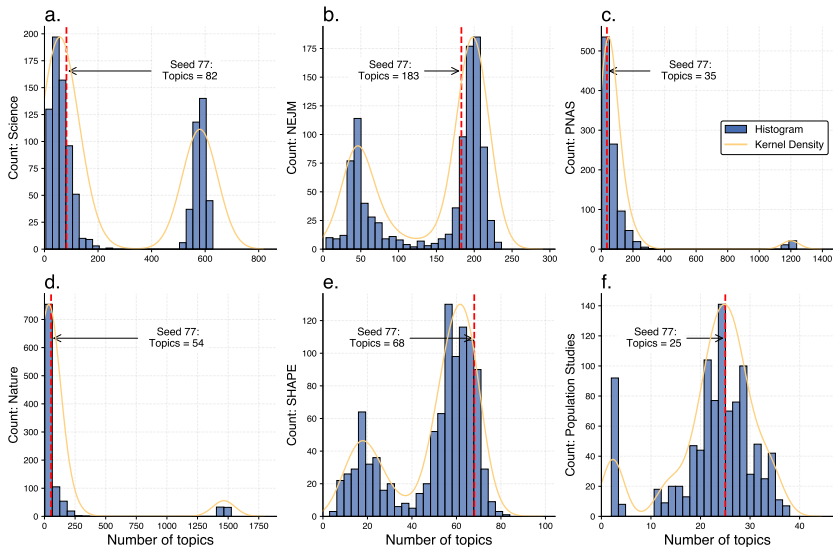
- **Problem Statement:** By setting one seed, we have no idea about the potential key variation of our estimand as a function of how random numbers were generated. This is computationally unintensive, but scientifically dour.

- **Problem Statement:** By setting one seed, we have no idea about the potential key variation of our estimand as a function of how random numbers were generated. This is computationally unintensive, but scientifically dour.
- **Solution:** Visualize the outcome space of a of large number (10k? 100k?) seeds simultaneously. This is computationally intensive, but scientifically illuminating.

- **Problem Statement:** By setting one seed, we have no idea about the potential key variation of our estimand as a function of how random numbers were generated. This is computationally unintensive, but scientifically dour.
- **Solution:** Visualize the outcome space of a of large number (10k? 100k?) seeds simultaneously. This is computationally intensive, but scientifically illuminating.
- **Replication:** We can consider whether an original result is in the tail of our distribution (or IQR?) or not.

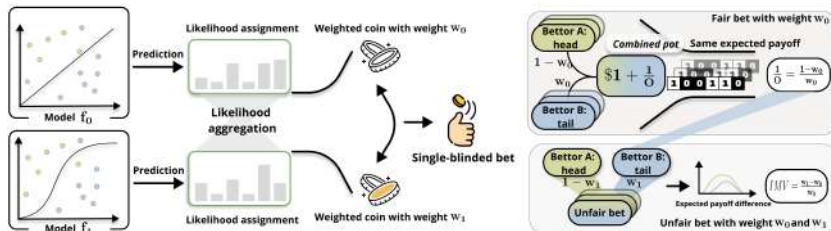


‘From a Seed of Doubt Grows a Forest of Uncertainty’
 With Arun Frey, Jiani Yan and Mark Verhagen



‘From a Seed of Doubt Grows a Forest of Uncertainty’
 With Arun Frey, Jiani Yan and Mark Verhagen

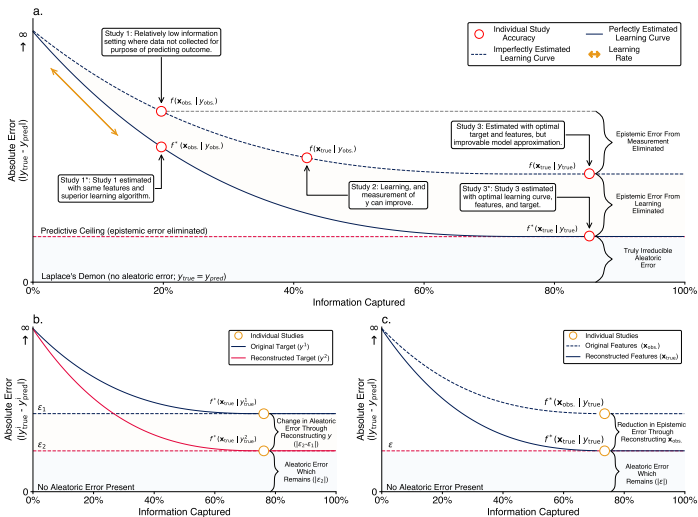
Introducing the InterModel Vigorish!



The InterModel Vigorish (IMV) is a dimensionless, log-likelihood based metric that quantifies the additional predictive “edge” of an enhanced model over a baseline.

1. Binary framework
2. IRT framework
3. SEM/CFA framework
4. ML/DL framework

Mathematical Philosophy of Prediction



Am I Determined to be Who I am?



"THEY FOUND YOUR GENE FOR ANTI-SOCIAL BEHAVIOR, BUT
THEY'RE STILL LOOKING FOR THE ONE FOR ROBBING LIQUOR STORES."

ABMs of Healthcare Demand and Supply

Incoming HDS CDT studentship. The project will:

- Design a working replica of the UK Healthcare System.
- Primary, Secondary, and Social Care.
- Calibrate this with 'Big' NHS data and UK census.
- Embed agents with the LLMs for decision making abilities.
- Match to moments of observed healthcare system.
- Ideally, make this public-facing.

Conclusion

I hope I've been able confer my excitement, enthusiasm and inspiration for:

- 'Assumption free' health and social science.
- 'Semi-structured' data found 'in the wild'.
- Openness and Transparency in (Care) research.
- An appreciation of how to better evaluate quantitative research, uncertainty, and the value of 'Openness'.

If just one more 'open' paper, program, design, or estimator enters scientific record, everything here will have been worth while it ☺.

Thank you for attending!

Thank you to my collaborators!

Don't hesitate to get in touch at:

charles.rahall@demography.ox.ac.uk

and

github.com/crahall!