



Unlocking the Potential of Internet Search Data for Health & Social Science Research

12 June 2025 Roundtable

Royal College of General Practitioners | 30 Euston Square, G.18 | London, UK

Agenda

June 12, 2025 Roundtable Workshop
Unlocking the Potential of Internet Search Data for Health & Social Science Research

Welcome + Introductions

10:00-10:35

Hosts: Rushil Ranchod, Chuy Chavez, Matthew Thompson

Session 1: Landscape + Opportunity

10:35-11:25

Moderator: Matthew Thompson

Panelists: Garth Funston, Richard Graham, Charles Marshall

Break

11:25-11:40

AM Tea & Coffee

Session 2: Tools + Governance

11:40-12:30

Moderator: Chuy Chavez

Panelists: Tom Fish, Jessica Bell

Break

12:30-13:15

Lunch

Session 3: Data Integration @ Scale

13:15-14:00

Moderator: David Zendle

Panelists: James Flanagan, Aiden Doherty, Luke Sloan, Anya Skatova

Session 4: Challenges for Research

14:00-14:45

Moderator: Richard Graham

Panelists: Suzanne Scott, Ali Connell, Tim Chico

Break

14:45-15:00

PM Tea & Coffee

Session 5: Looking Ahead

15:00-15:50

Facilitated Brainstorm Session

Moderators: Matthew Thompson & Eboney White

Break

15:50-16:00

Transition

Session 6: Hybrid Closing + Wrap Up

16:00-17:00

Hosts: Rushil Ranchod, Chuy Chavez, Matthew Thompson

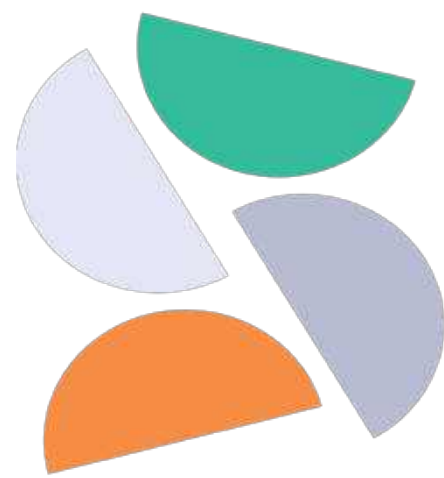
Reflections: Emmanouil Tranos & Agnieszka Scott

June 12 Round Table Presentation

Table of Contents

Session	Link to Slides	Moderator/Panelist
Welcome + Introductions	Smart Data Research UK	Dr. Rushil Ranchod, PhD
Session 1: Landscape + Opportunities	Current State	Dr. Matthew Thompson
	Early Cancer Detection	Dr. Garth Funston, PhD
	Early Gynaecological Malignancy	Dr. Jen Barcroft
Session 2: Tools + Governance	Data Portability Individual Data	Chuy Chavez
	Data Portability	Tom Fish
	Ethical & Legal Considerations	Dr. Jessica Bell, PhD
Session 3: Data Integration @ Scale	Data Scaling Access	Dr. David Zendle, PhD
	Data Integration Wearables Data	Dr. Aiden Doherty, PhD
	Digital Trace Data (no slides)	Dr. Luke Sloan, PhD
	Loyalty Card Data (no slides)	Dr. James Flanagan, PhD
	Digital Footprint Lab	Dr. Anya Skatova, PhD
Session 4: Challenges for Research	Public Awareness Consent	Dr. Suzanne Scott
	Cohort Studies (no slides)	Dr. Ali Connell
	Study Design (no slides)	Dr. Tim Chico

Welcome + Introductions



**Smart
Data
Research
UK**

Dr. Rushil Ranchod, PhD
Senior Manager

Smart Data Research UK

The UK's National Programme for Smart Data Research

Rushil Ranchod, PhD

Senior Manager: Research Strategy and Impact



UK Research
and Innovation

Let's do good things with data



Partnerships

- Building partnerships across government, funders, industry, academia to support SD access and use

Innovation and impact

- 9 accelerator awards to pilot new data sources, develop new tools and methods to help solve social and economic challenges.
- Stimulating smart data research – new funding opps

Public engagement

- UK-wide public deliberations to better understand how people feel about smart data research.
- Dialogue Report is available on SDRUK website

Legal and ethics resource

- Commissioned research to baseline E&L frameworks and principles on different types of SD
- Identify opportunities for SDR UK to fulfill in this space

Contact Smart Data Research UK



rushil.ranchod@esrc.ukri.org



smartdataresesarch@ukri.org



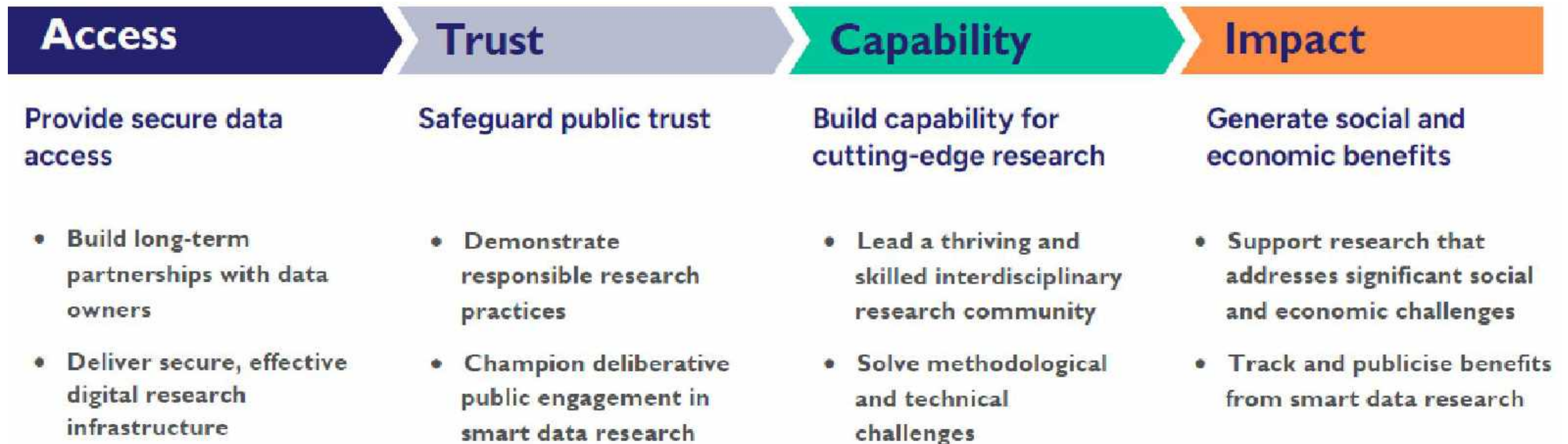
Smart Data Research UK



www.sdruk.ukri.org



SDR UK is a UKRI **infrastructure investment** which seeks to harness the power of digitally derived data to drive research and innovation for public good



SDR UK is a UKRI *infrastructure* investment.

We fund a portfolio of national data services to enable better access to smart data

Acquire, steward and enable safe access to smart data for research:

Partnerships with data owners; developing and curating data products; address issues of representativeness, provenance, bias, licensing; protect sensitive data when making it available

Collaborate to build a user-friendly federation of services and enable cross-domain research:

SDR data will be FAIR; collaboration with other UK data services to standardise and streamline services

Ensure responsible use of data:

Go beyond regulatory compliance to embed ethics and responsibility in all aspect of the data service

Build capability:

Establish CoP; support development of research skills to build user base for smart data

Centres of excellence for smart data research:

Impact-focused research to shift practice, thinking and capacity; bring researchers, policymakers and other actors together to address critical challenges in the UK



A network of expert data services, with central coordination

Smart Data Research Data Services



Data Service	Lead Organisations	SDR UK Thematic Pillars	Data sources
Smart Data Donation Service	University of York	Digital Society	Video game (donation)
Geographic Data Service	UCL University of Liverpool	Productivity and Prosperity	Web/App and Image, Financial, Transport
Healthy and Sustainable Places Data Service (HASP)	University of Leeds	Health and wellbeing Sustainability	Mobility, Spatial, Housing Financial, Food, Health (NHS, Gyms, etc)
Imagery Data Service (Imago)	University of Liverpool Newcastle University	Sustainability Productivity and Prosperity Health and Wellbeing	Satellite, Imagery
Financial Data Service (FinDS)	University of Edinburgh Smart Data Foundry	Productivity and Prosperity	Finance data (current account, credit, insurance, etc)
Smart Energy Data Service (SENSE)	University of Oxford Energy Systems Catapult	Sustainability	Infrastructure data (smart meter data, EV charging, etc.)





Chuy Chavez
Engineering Manager
Google Takeout & Data Portability



Google for Health

Dr. Matthew Thompson
Clinical Research Lead

Housekeeping & Introductions

Session 1: Landscape + Opportunity

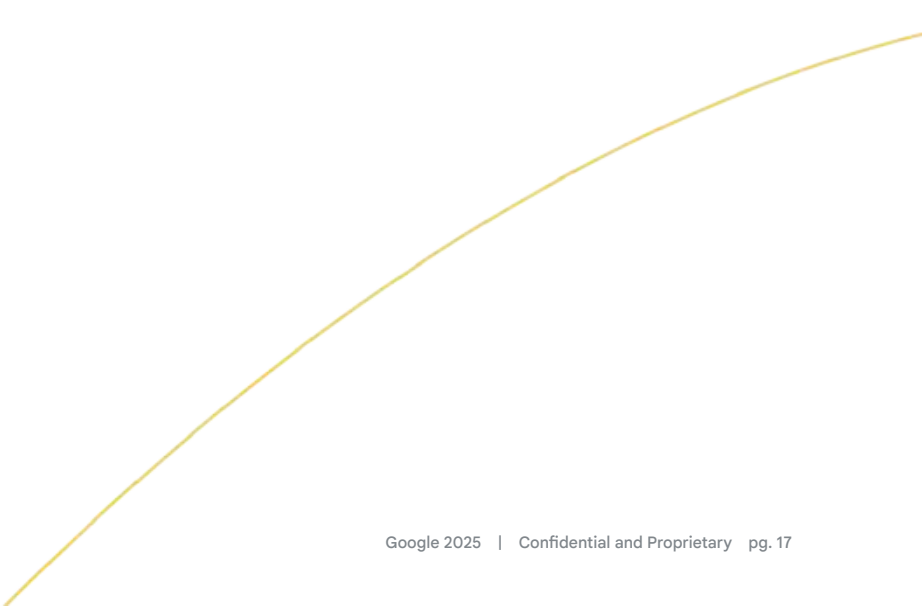
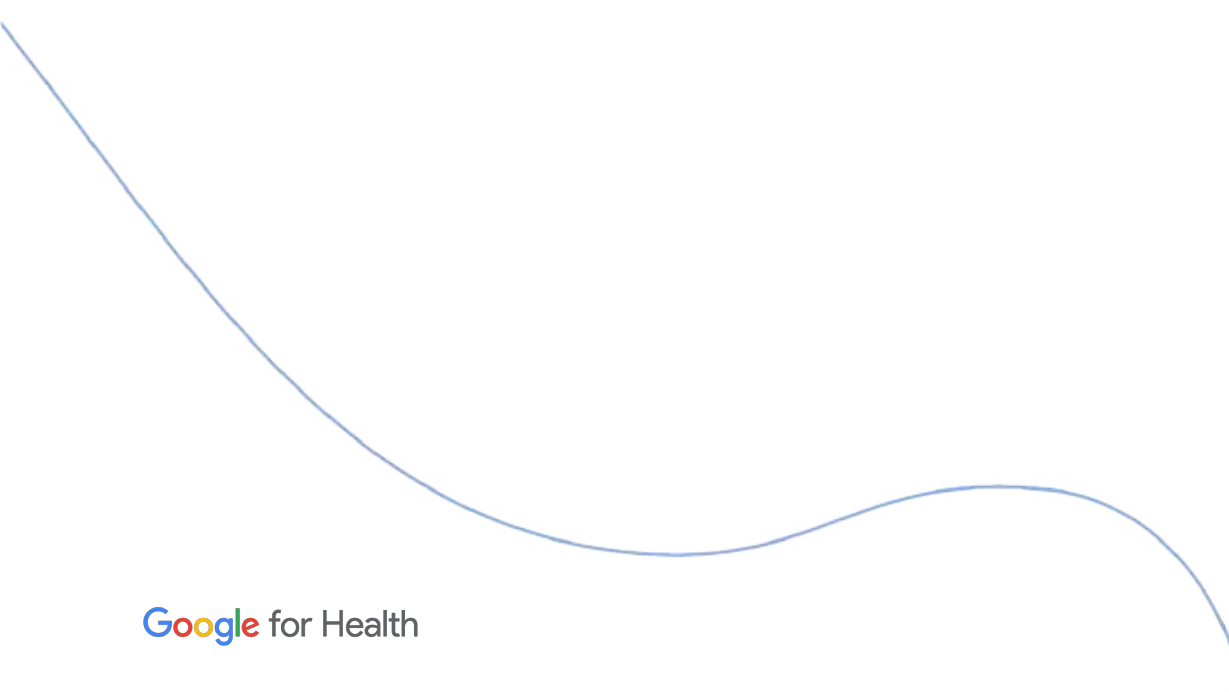


Dr. Matthew Thompson

Clinical Research Lead, Google for Health



“What would it take to publish 1000 studies
using internet search data?”



What studies have been published to date?

Systematic review of health research using individual level internet search data.

Thompson M et al. NPJ Digital Medicine *under review* <https://www.researchsquare.com/article/rs-4456499/v1>

23 studies used internet search data for diagnosis or prognosis research

- Data used
 - **Anonymous** search queries from backend database- Bing (9), Yahoo! (1)
 - **Consented** participants' searches -Google (9)
 - Both Bing and Google (1), not specified (2)
- Most from US, one from UK
- Number of participants 20 to 11,050
- Health conditions
 - Mental health (e.g. schizophrenia, suicidality, mood disorders)
 - Neurological (e.g. Parkinson's, ALS, stroke)
 - Malignancies (e.g. pancreatic, lung, gyn)

Notable features of studies using Google Takeout for health research

Populations, settings

- mostly USA studies (one from UK)
- Participants younger, higher risk (perhaps appropriately) recruited usually from academic centers
- Less representative of broader / at risk populations

Recruitment

- Surprisingly high consent rates (approx 50%)
- Significant proportions excluded due to lack of Google account or technical difficulties
- Some (not much) qualitative assessment of privacy/ethical concerns

Notable features of studies using Google Takeout for health research

Study type/methods

- Consented participants. Search data linked to EHR or surveys
- Small-sized, cohort studies
- Few control populations
 - Some compared different time periods for individuals (pre-post)

Predictors used from internet search data

- Temporal: numbers of searches, time of day/week, changes over time
- Linguistic: search terms labelled and aggregated using different methods: Informal coding, NLP, semantic methods

Analytic issues

- Multiple techniques used from more to less complex
- Small sample sizes/underpowered
- Multiple comparisons
- Multiple potential confounders

The Opportunity

Emerging Use Cases

**General Population &
Birth Cohorts**

**Disease-Specific &
Aging Cohorts**

Data Integration Pilots

**Social Sciences &
Mental Health**

**Retrospective v.
Prospective**

Micro Data Linkages

Could internet search data support early cancer detection?



Dr. Garth Funston, PhD

Senior Clinical Lecturer in Primary Care Cancer Research

Wolfson Institute of Population Health

Queen Mary University of London, UK

Email: g.funston@qmul.ac.uk

□ Patient interval: symptom onset - presentation



- Symptom appraisal
- Information seeking
(*E.g. internet search*)
- Self-manage
- Seek support

□ Current evidence ?

PARTIAL ACCESS | Original Contributions | June 07, 2016



Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results

Authors: [John Paparrizos, MSc](#), [Ryen W. White, PhD](#) , and [Eric Horvitz, MD, PhD](#) | [AUTHORS INFO & AFFILIATIONS](#)

Brief Report

FREE

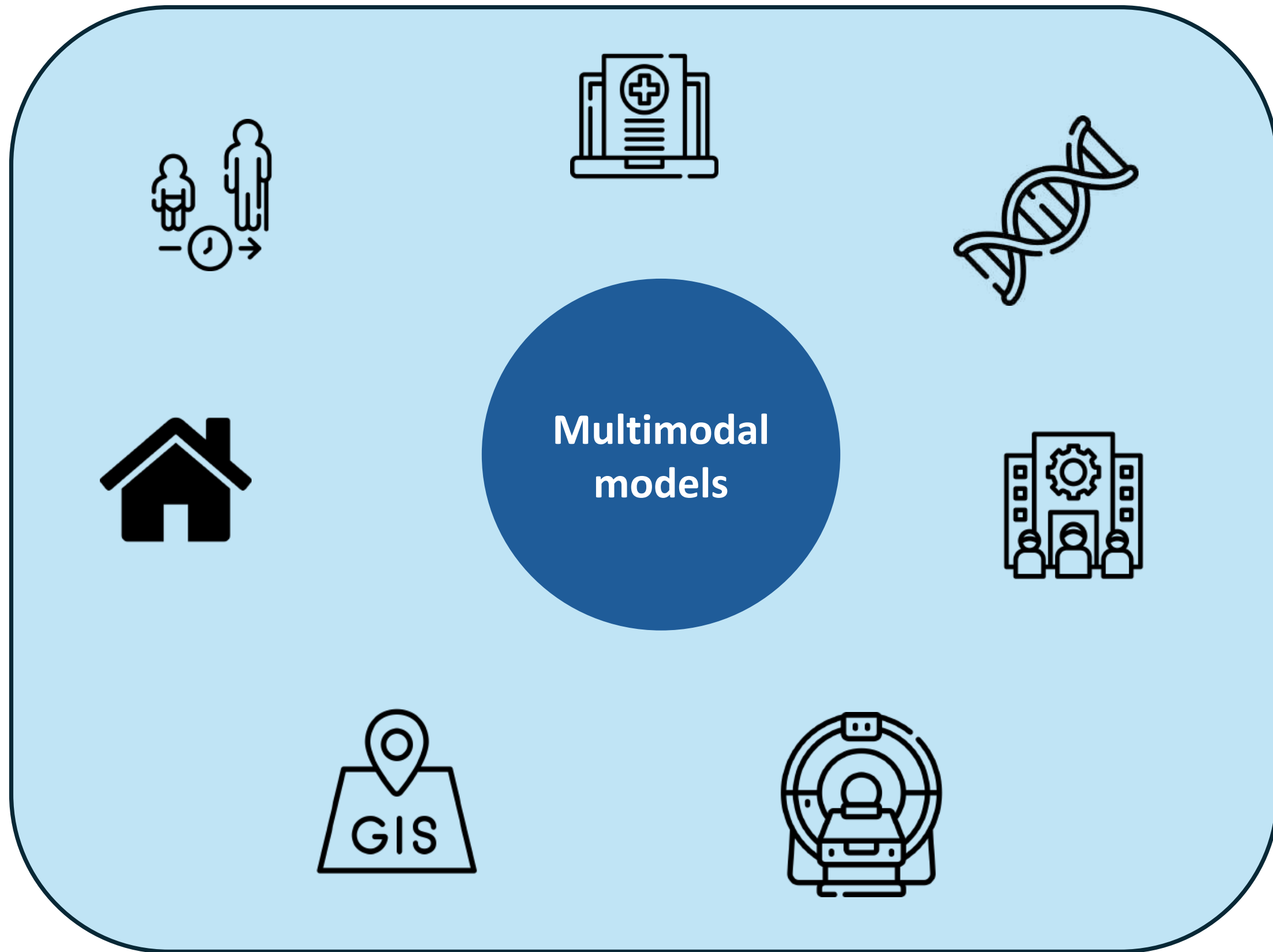
Evaluation of the Feasibility of Screening Patients for Early Signs of Lung Carcinoma in Web Search Logs

Ryen W. White, PhD¹; Eric Horvitz, MD, PhD¹

Research | [Open access](#) | Published: 11 March 2024

Using online search activity for earlier detection of gynaecological malignancy

[Jennifer F. Barcroft](#), [Elad Yom-Tov](#), [Vasileios Lampos](#), [Laura Burney Ellis](#), [David Guzman](#), [Víctor Ponce-López](#), [Tom Bourne](#), [Ingemar J. Cox](#) & [Srdjan Saso](#) 



Using online search activity for earlier detection of gynaecological malignancy

Dr. Jen Barcroft MBChB MRCOG, PhD

Clinical Research Fellow

Obstetrics & Gynaecology | Early Detection

Imperial College London

BACKGROUND

Gynaecological cancer
Ovarian- most lethal (1 in 56)¹
Endometrial- most common (1 in 39)²

No screening program in place

Symptom detection

Patient recognition
Triage by primary care

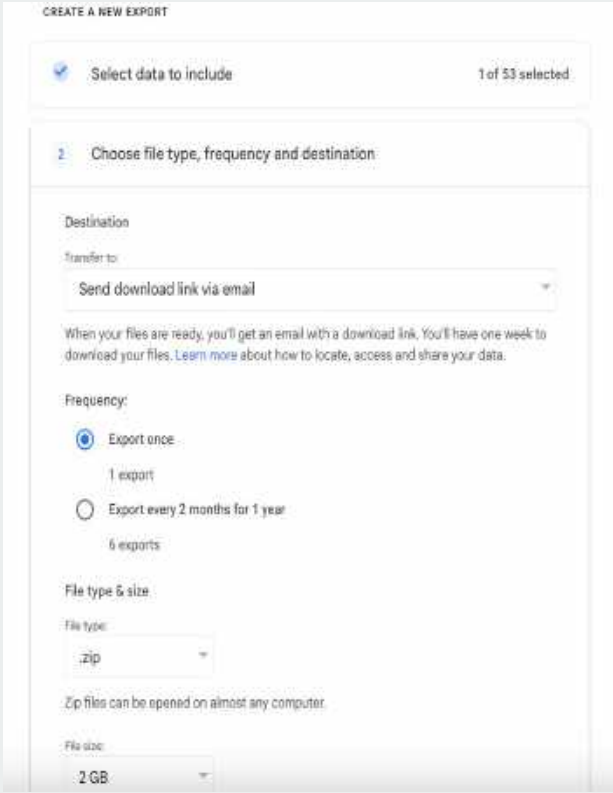
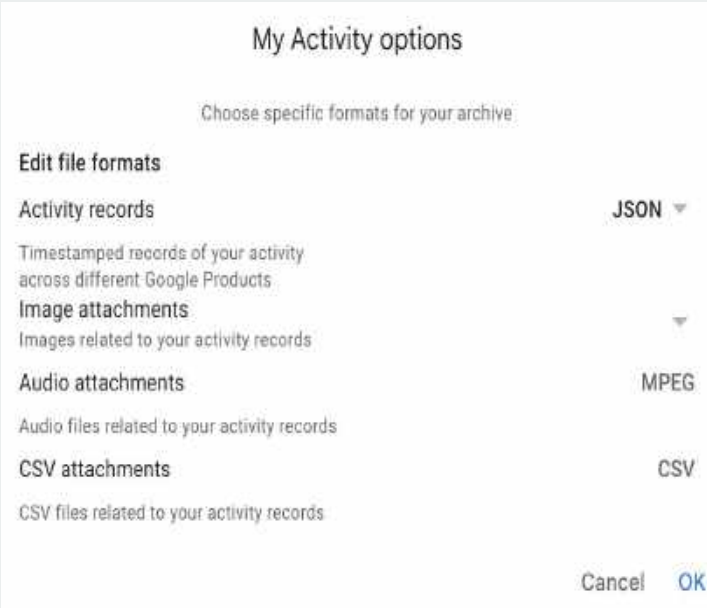
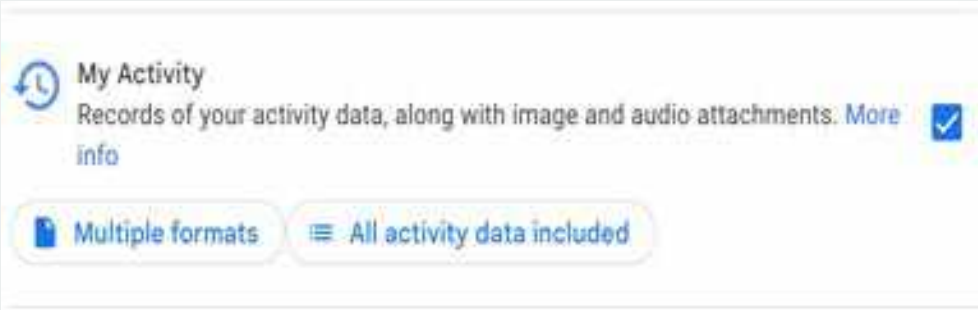
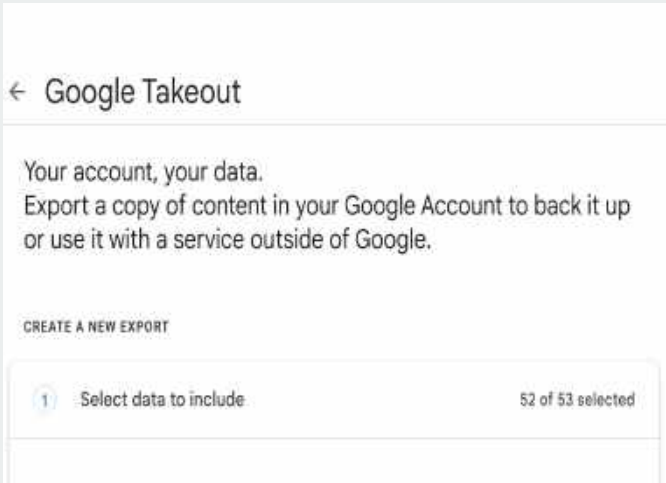
Primary care

imaging
Ultrasound

Objectives

1. Evaluate whether **online search patterns are different in women with a benign or malignant diagnosis**
2. Can they be used to **enable the earlier detection of Gynaecological cancer?**

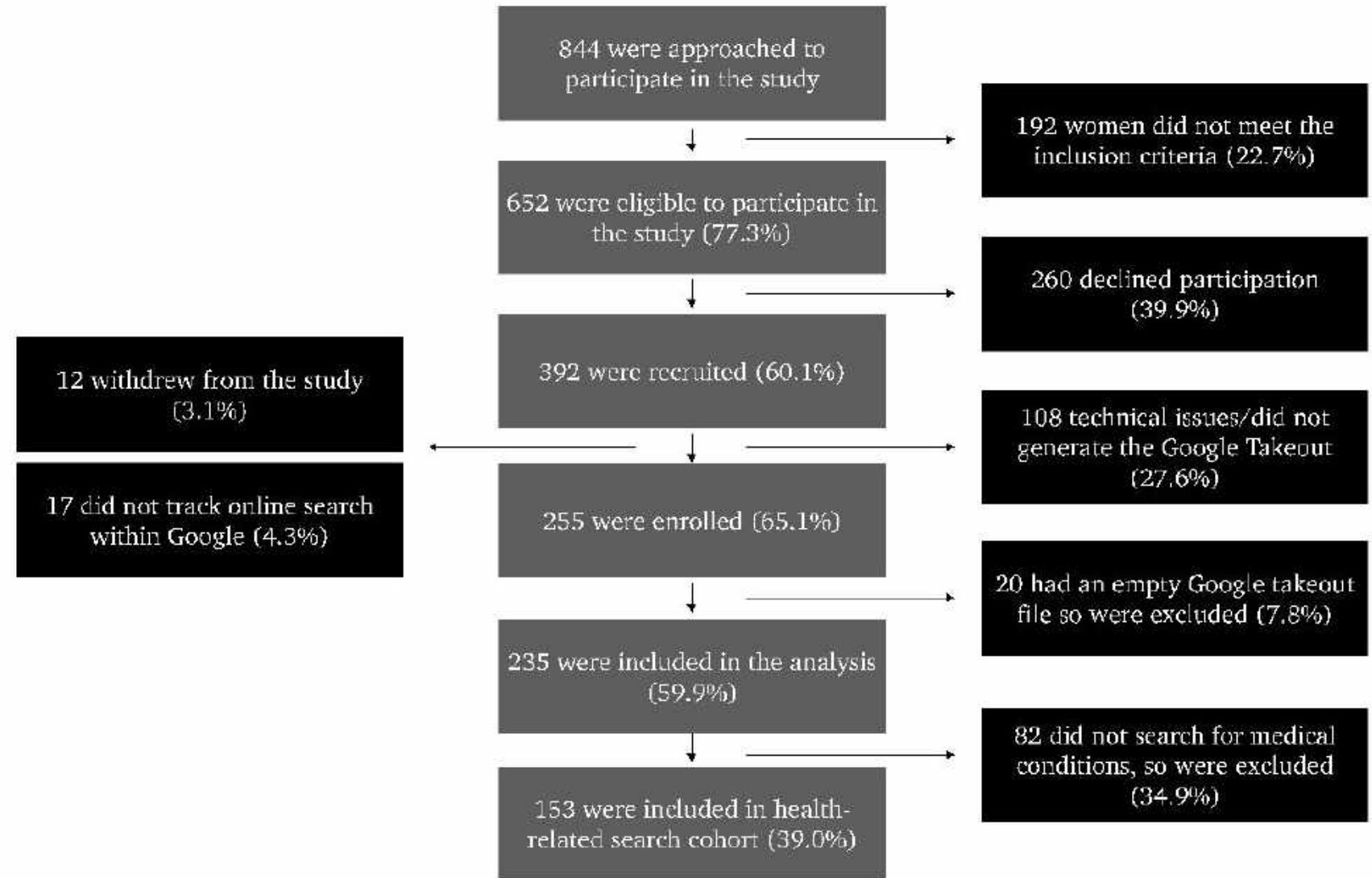
Google takeout pipeline



Clinical recruitment

Women (≥ 18 years old)
with:

1. **Gynaecological symptoms** referred on an **urgent cancer pathway** to secondary care
2. **GOOGLE** account



Clinical data

Clinical Questionnaire (page 1)

Demographics and Medical history

Age | Height | Weight | Ethnicity

Do you have any of the following conditions? If yes, please circle:

Diabetes mellitus | Low or High thyroid hormone (hyper/hypothyroidism)
Asthma | Autoimmune disorders | High blood pressure (hypertension)

Have you been diagnosed with covid-19 infection? (yes/no) If yes, specify the date:

Gynaecological history

When did your periods first start (Menarche)?

Are you known to have any of the following conditions, if so when were you diagnosed?

Endometriosis | Uterine Fibroids | Polycystic Ovarian Syndrome | Adenomyosis

Have you had any previous pelvic surgery? (please circle all that apply)

Uterine polyps | Fibroids | Ovarian Cysts | Endometriosis | Tubal

Uterine Septum | Hysterectomy (removal womb) | Oophorectomy (remove ovary)

Do you have a history of pelvic infection (i.e., Chlamydia/Gonorrhoea)? (yes/no)

Have you undergone any gynaecological investigations? (yes/no)

Pelvic ultrasound | Endocrine screen | Tumour markers | Diagnostic laparoscopy

Pre-Menopausal:

Do you currently use any contraception or have done in the last year?

Combined pill | Progestogen only pill | Natural cycle | Mirena Coil

Nexplanon | Copper IUD | Condoms

Do you have regular periods? (yes/no)

How long is your usual menstrual cycle (cycle start = first day of period)?

How many days does your period normally last for?

Have your periods currently stopped for more than 6 months (Amenorrhoea)? (yes/no)

Do you have to take time off work or use regular pain killers due to your period? (yes/no)

Have you experienced any of the following symptoms? (please circle all that apply)

Excess hair (face, arms) | Acne | Loss of hair (head)

Fertility

Are you currently wishing to conceive? (yes/no)

Do you have any history of subfertility? (yes/no), if so how long have you been trying to conceive?

Have you undergone fertility treatment previously? (yes/no)

Have you undergone treatment for recurrent miscarriage? (please circle all that apply)

Progesterone | Aspirin | Steroids | Low Molecular Weight Heparin

Have you undergone any complementary therapies (acupuncture, reflexology)?

Has your partner undergone semen analysis? (yes/no)

if yes, was the result: Normal| Abnormal (specify abnormality:)

Have you experienced the following symptoms in the last 12 months?	Y/N	When did the symptoms start?	How often do you experience the symptoms? (daily, 1-5/week, fortnightly, monthly)	At their worst, how severe are your symptoms? (1-10 with 10 being the most severe)
Pelvic pain				
Bleeding after sexual intercourse				
Increased abdominal size (girth)				
Abdominal bloating				
Appetite loss				
Constipation				
Diarrhoea				
Sudden surge to pass urine (urgency)				
Passing urine more frequently (frequency)				
Passing urine at night (nocturia)				
Weight loss				
Weight gain				
Feeling full (early satiety)				
Reflux				
Pain during sex (Dyspareunia)				
Change in vaginal discharge				
If pre-menopausal, do you experience:				
Heavy periods				
Bleeding in between periods				
Painful periods (dysmenorrhea)				
Pain when opening bowels (dyschezia)				
If post-menopausal, do you experience:				
Bleeding after the menopause				

235 symptomatic women (median age 53, range 20-81)

Gynaecological diagnosis

benign n= 174 (74.0%), malignant n=61 (26.0%)

Google Takeout export
(online search history)

Clinical Questionnaire

Health filter applied

Filtered Google Takeout
(24 months) and
pseudo-anonymised

**Model
Development
Vector Space**

Search terms

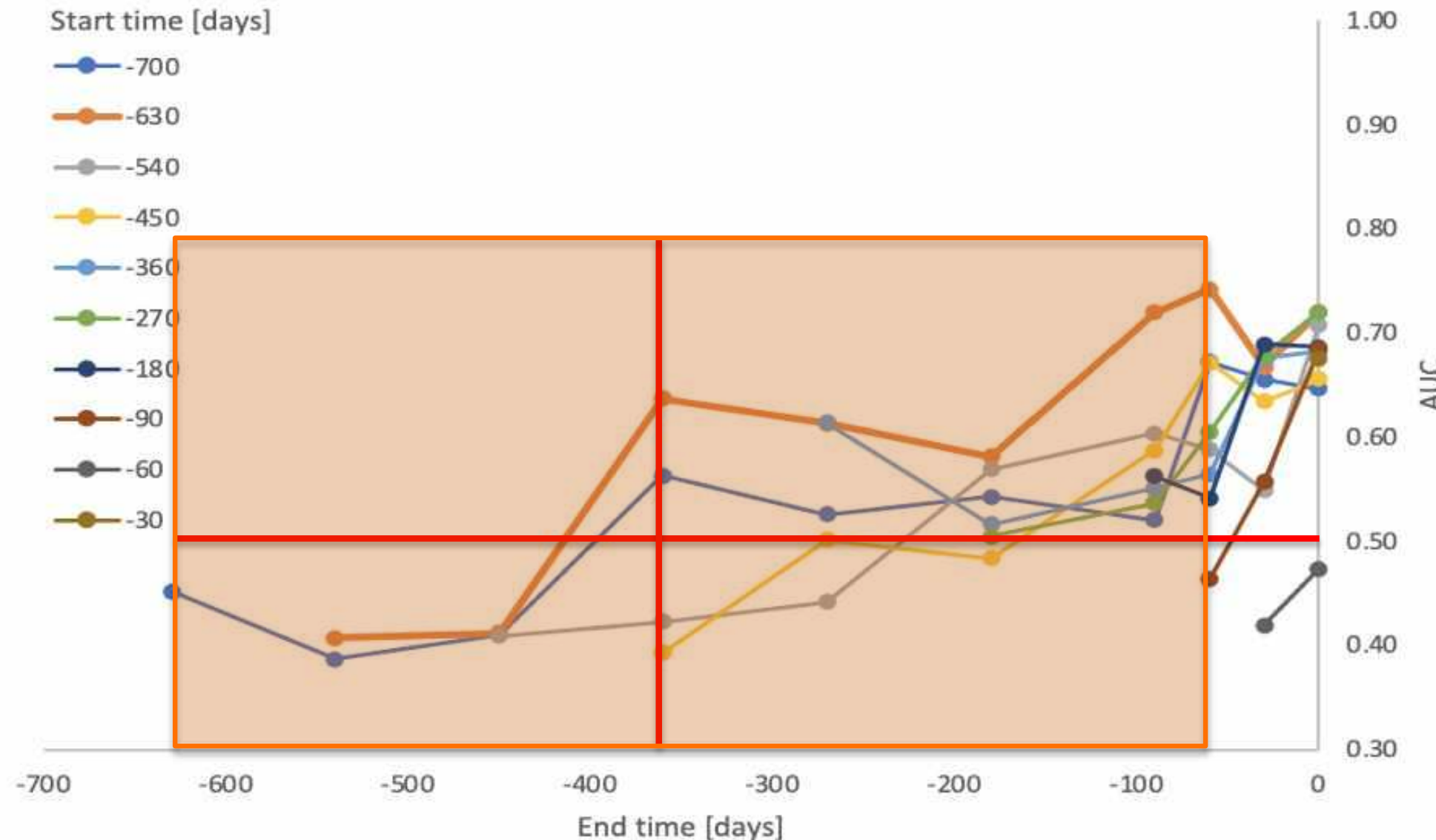
Search terms separated into
health categories

Terms model (all
terms/word pairs by
five individuals)

Categories model

Questionnaire
model

Online search patterns appear different in benign and malignant gynaecological conditions

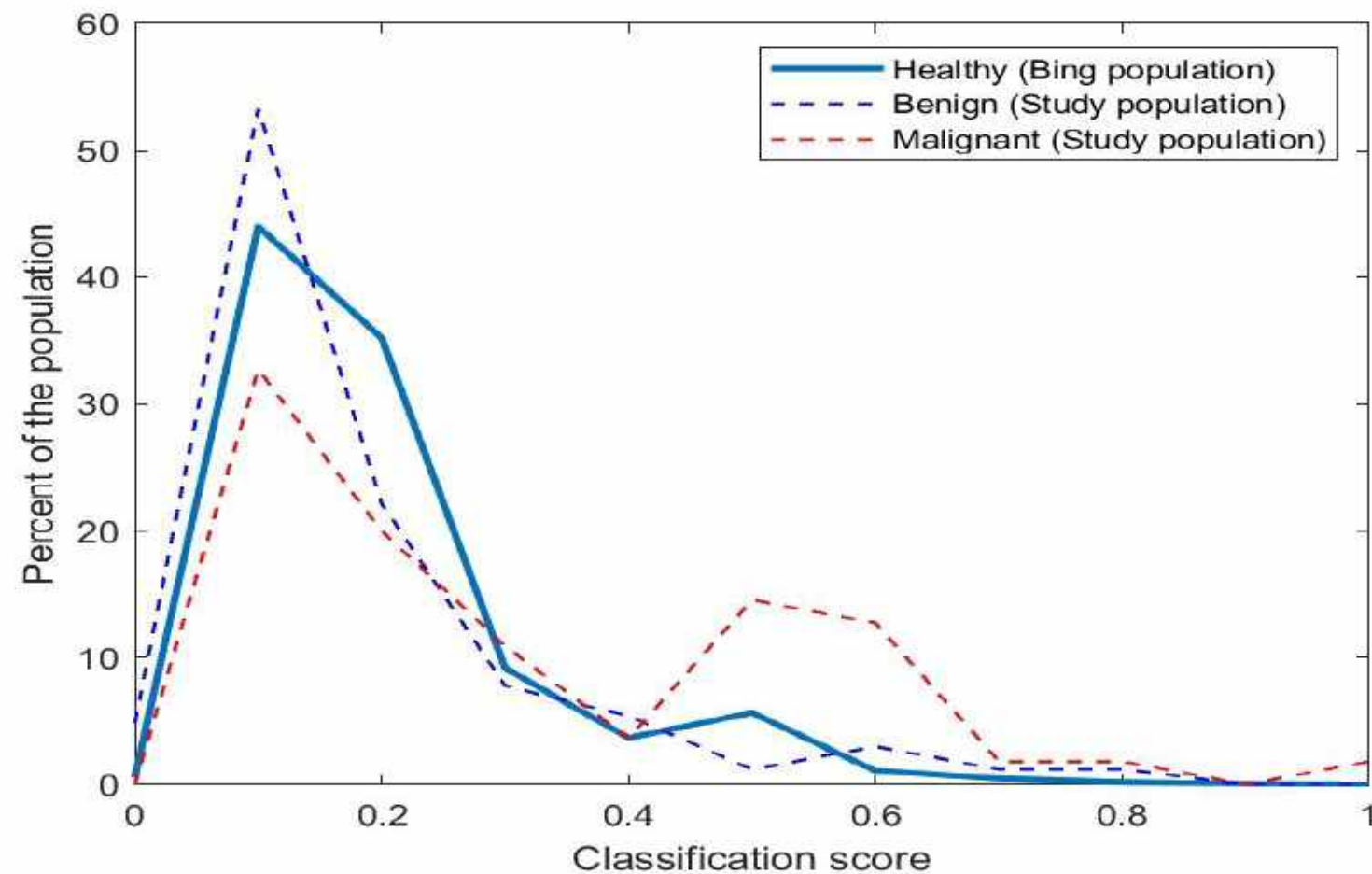


- Search terms model
- Difference in search terms between benign and malignant first seen **360 days before GP referral (AUC 0.64)**
- Best performing model (630-60 days) seen **60 days in advance of GP referral (AUC 0.74)**
- Excluding individuals (n=82) who did not search for health terms **improved the performance** of the predictive model (**AUC 0.82**) in health search cohort (n=153)

Bing Control population

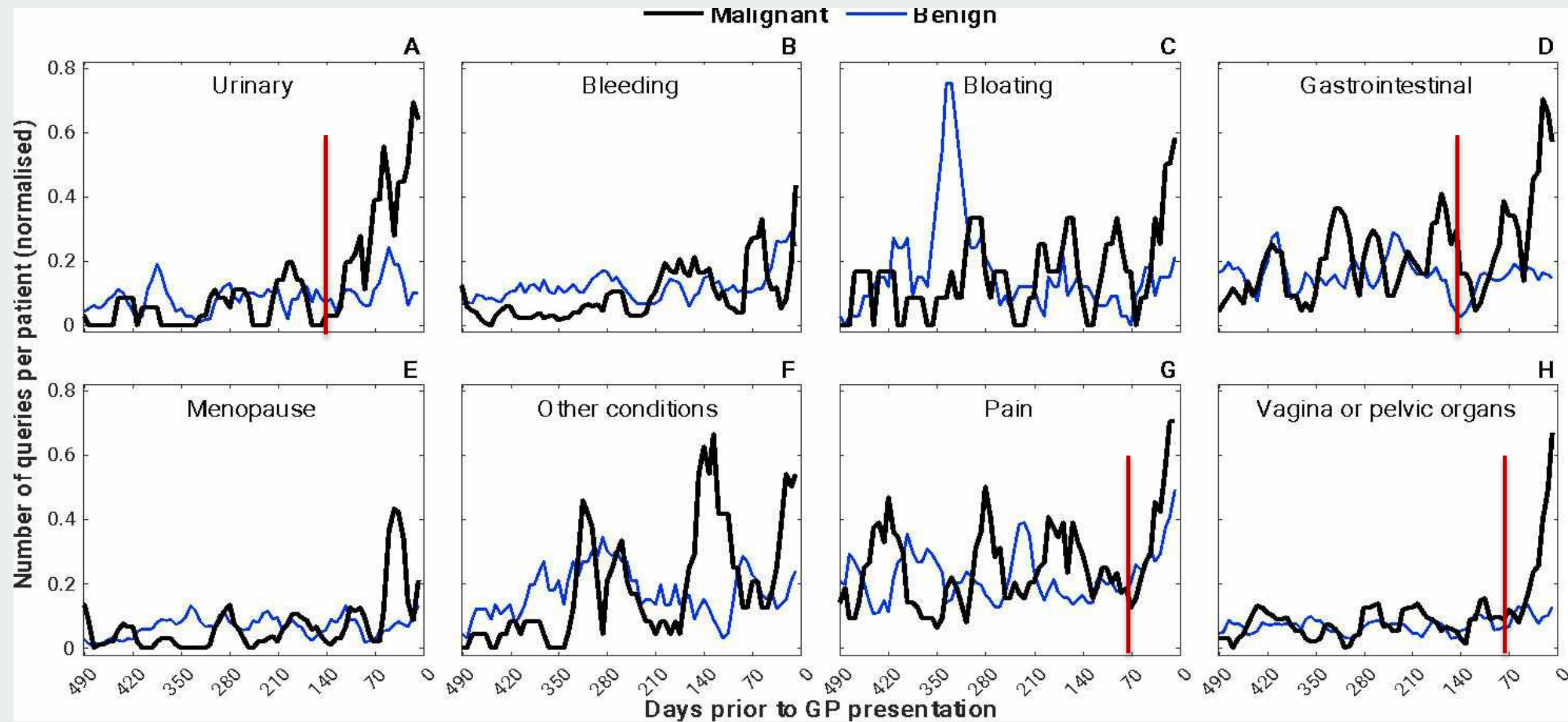
- **1.8 million users** searched for at least **one keyword** from the medical key word list AND made **one search query each month** between October 2021-September 2022
- Female and male (anonymous)
- Individuals searched for **gynaecological cancer ten or more times** during the data period (Oct 2021-June 2022) or three months after (July-Sept 2022) were excluded
- Population- searched for terms **related to gynaecological cancer**, but **assumed not** to have an active gynaecological diagnosis

Search terms of Bing control group mirrors benign group



- Model ($T_1 = -270$ and $T_2 = -1$) trained on data from all participants- benign and malignant cases) was applied to search queries made in the data period (Oct 2021-June 2022) by Bing control group.
- Bing population typically **low classification scores** suggesting the **absence of malignancy**

Different search patterns in benign and malignant gynaecological conditions



Summary



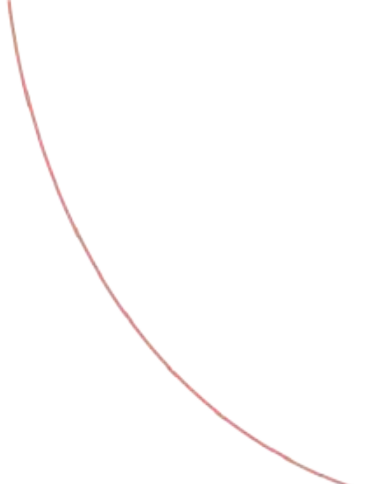
Feasible to use online search data in clinical research



Online search data may have **value in early signals of disease**



Large dataset required to evaluate this association



Dr. Richard Graham

Mental Health | Inclusion



Smart
Data
Donation
Service



UNIVERSITY
of York



UK Research
and Innovation

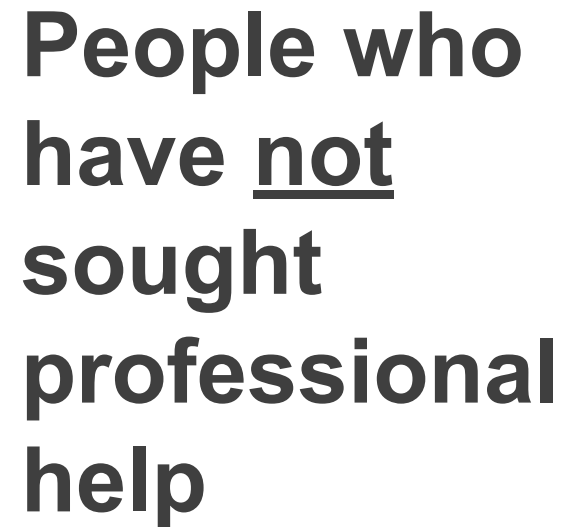
Leveraging Search Data To Develop Relevant Mental Health Support

Dr Richard Graham
Consultant Psychiatrist



- ❑ 2007: 25% Adults have a Mental Disorder; 75% of them had no support
- ❑ Qualitative Research enriched by Search Research and Social Listening
- ❑ Research informed development of an Online Service; Digital Marketing, led to Relevant, Relatable Supports
- ❑ Collaboration between Public Health, Clinical, Behavioural Science and Digital Marketing Professionals; full potential remains untapped.
- ❑ Community and faith leaders also key to uptake.

Device	Search Query	Query Type	Clicks	Sessions
Mobile	i can t sleep all i think about is horrible things	Can't Stop Thinking	1	1
Mobile	can t sleep due to cough	Cough	1	0
Mobile	cant sleep feel hungry	Hungry	1	0
Mobile	i cant sleep when i have work	Work	1	0
Mobile	islam reason why i cant sleep	Why	1	1
Mobile	why people cant sleep at night its feel like nigh	Why	1	0
Mobile	can t sleep and not slept	Not Slept	1	0
Mobile	currently 1 am ans i cant sleep	Time	1	1
Mobile	reddit can t sleep	Reddit	1	1
Mobile	can t eat and can t sleep and have numb hand	Symptoms	1	1
Mobile	cant aleep	Can't Sleep	1	1
Mobile	when u cant sleep even though you had hypnc	Hypnosis	1	0
Mobile	can t sleep waking up because of period cramp	Period	1	0
Mobile	i can t sleep because my mind keeps racing	Mind Racing	1	1
Mobile	i m at a sleepover and i can t sleep	Sleepover	1	0
Tablet	can t sleep anxiety stomach	Anxiety	1	1
Tablet	what should u do when u cant sleep	What To Do	1	0
Desktop	can t sleep test tomorrow	Exams	1	0
Mobile	36 weeks pregnant and can t sleep	Pregnant	1	1
Mobile	active brain cant sleep	Active Brain	1	1
Mobile	feeling tired but can t sleep	Tired But Can't Sleep	1	2
Mobile	flu and can t sleep	Flu	1	0
Mobile	smoking heroin for 8 months now can t sleep	Drugs	1	1



People who have sought professional help





NHS London / South London and Maudsley NHS
Foundation Trust

London Major Terror Incidents Search Research: Insights and recommendations

January 2019

Early Detection of Dementia

Professor Charles Marshall

QMUL Centre for Preventive Neurology

Honorary Consultant Neurologist

Barts Health NHS Trust and East London Foundation Trust

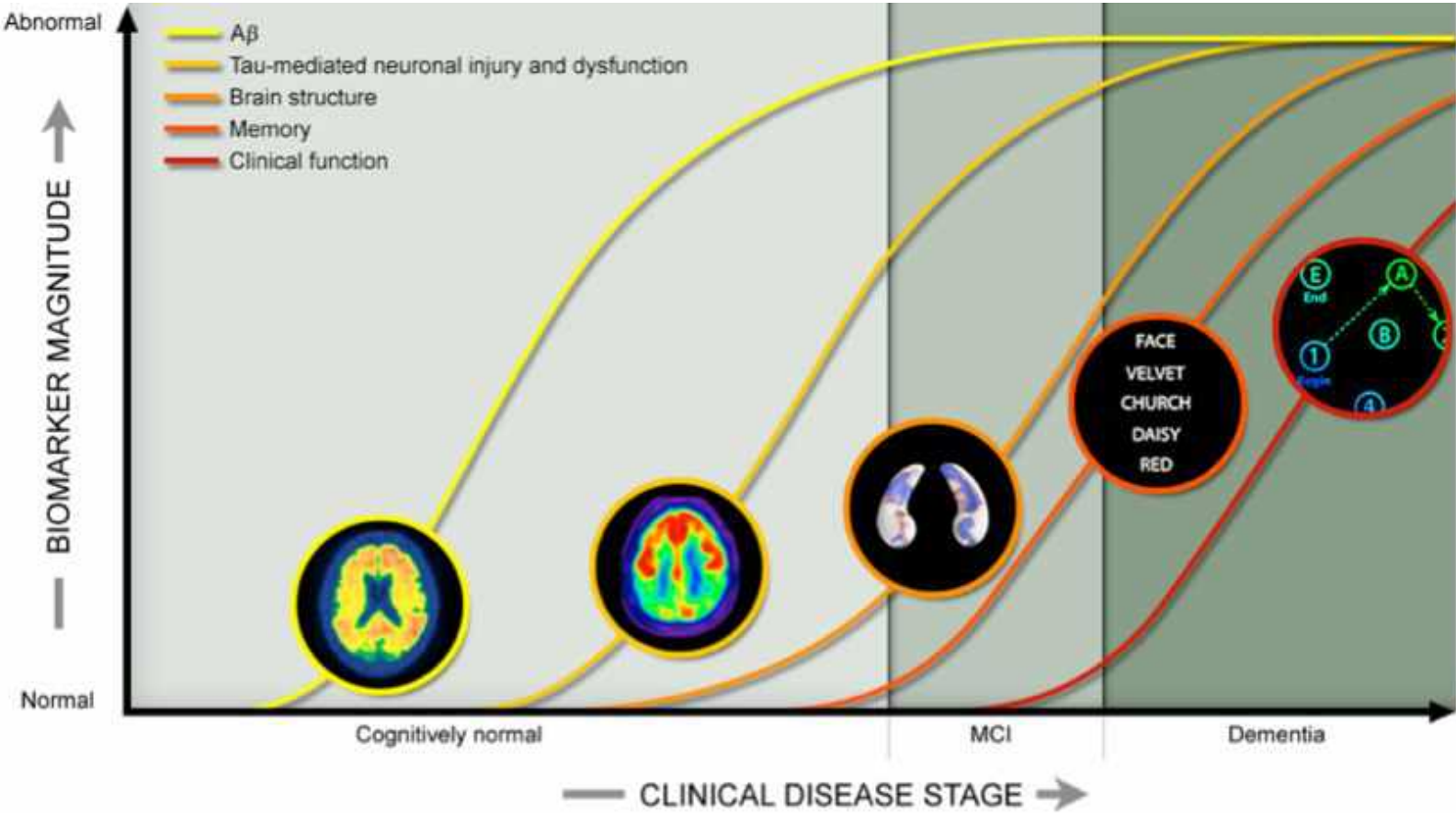
NHS London Clinical Director for Dementia

August 2018

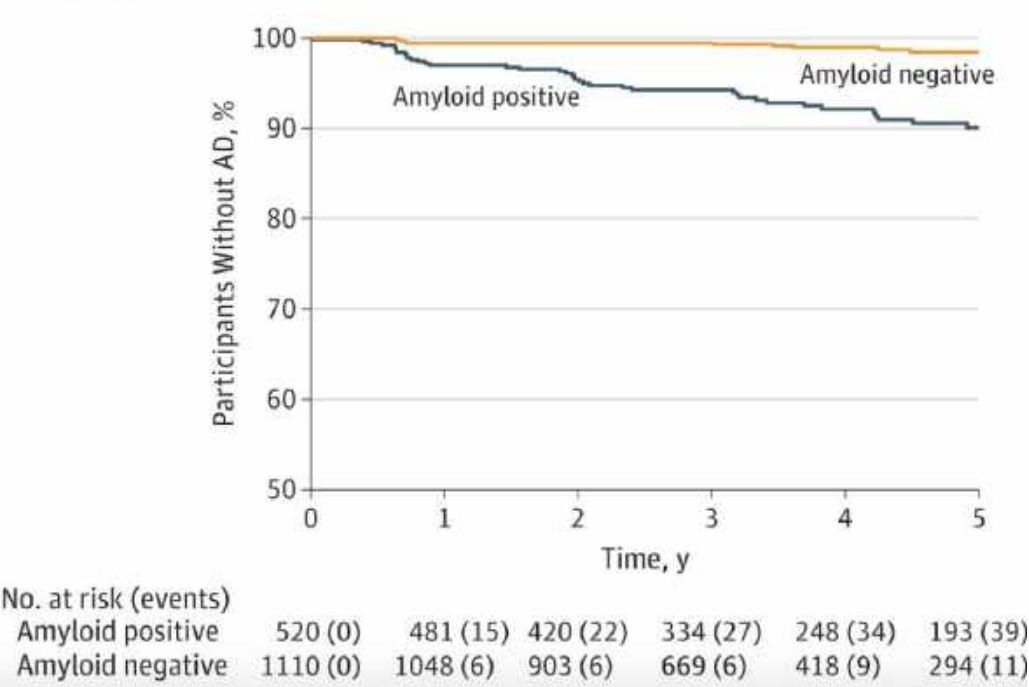
Prevalence and Outcomes of Amyloid Positivity Among Persons Without Dementia in a Longitudinal, Population-Based Setting

Rosebud O. Roberts, MB, ChB, MS^{1,2}; Jeremiah A. Aakre, MPH¹; Walter K. Kremers, PhD¹; [et al](#)

[» Author Affiliations](#) | [Article Information](#)



D Percentage of participants with no cognitive impairment or aMCI without AD dementia over time

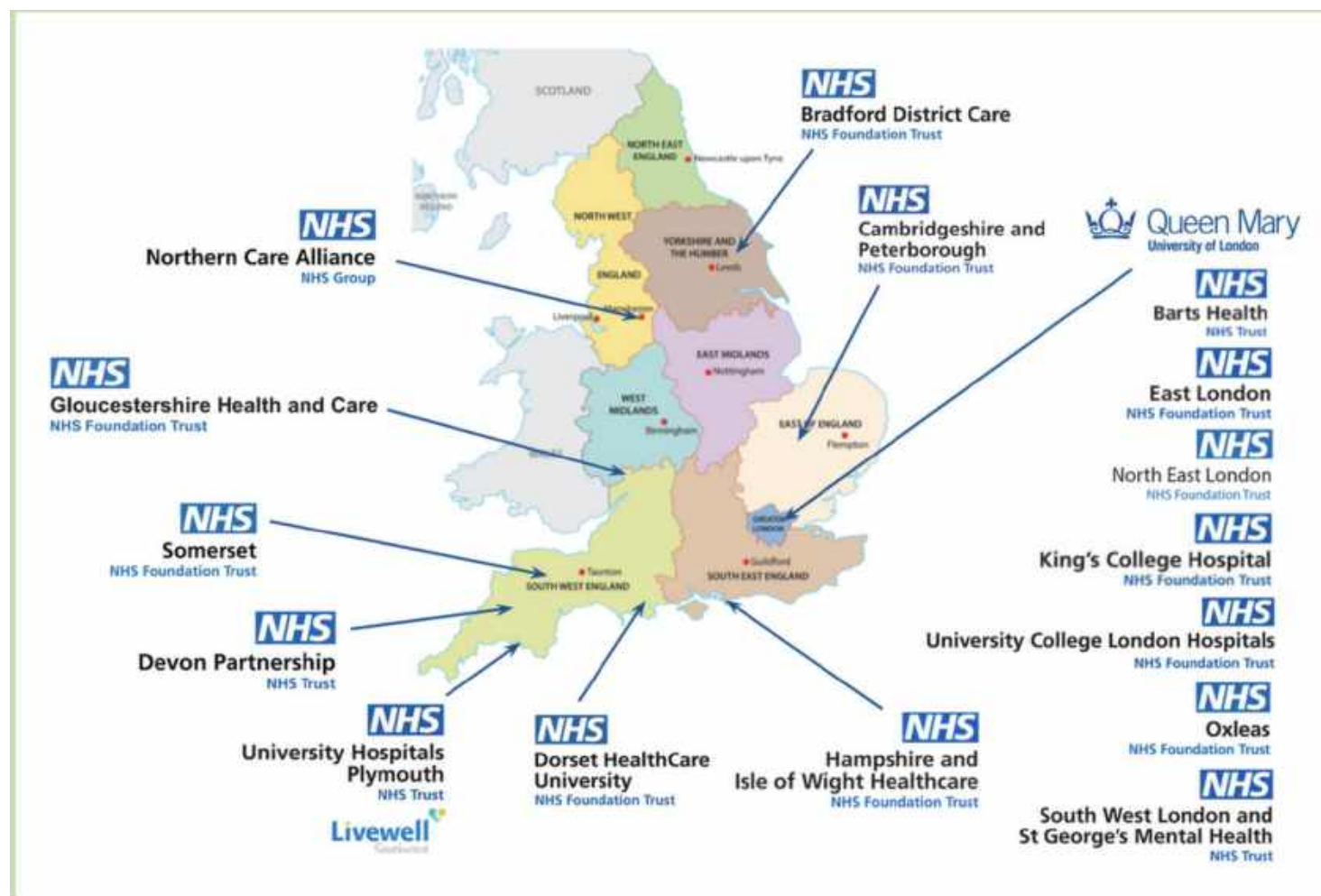




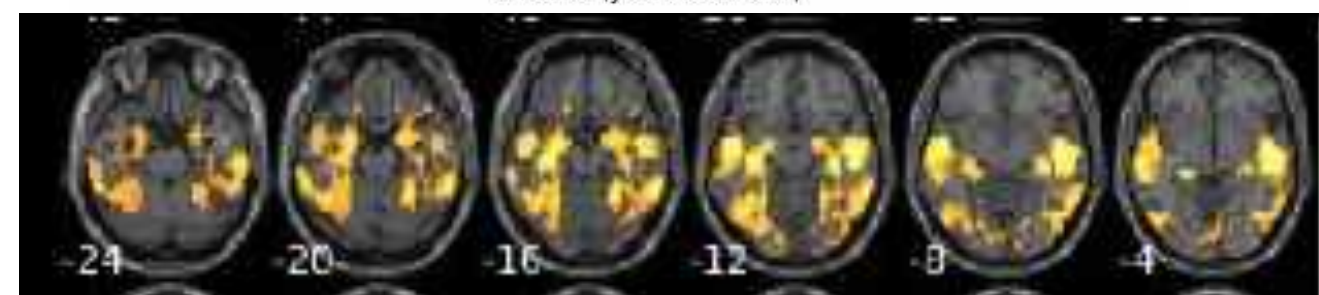
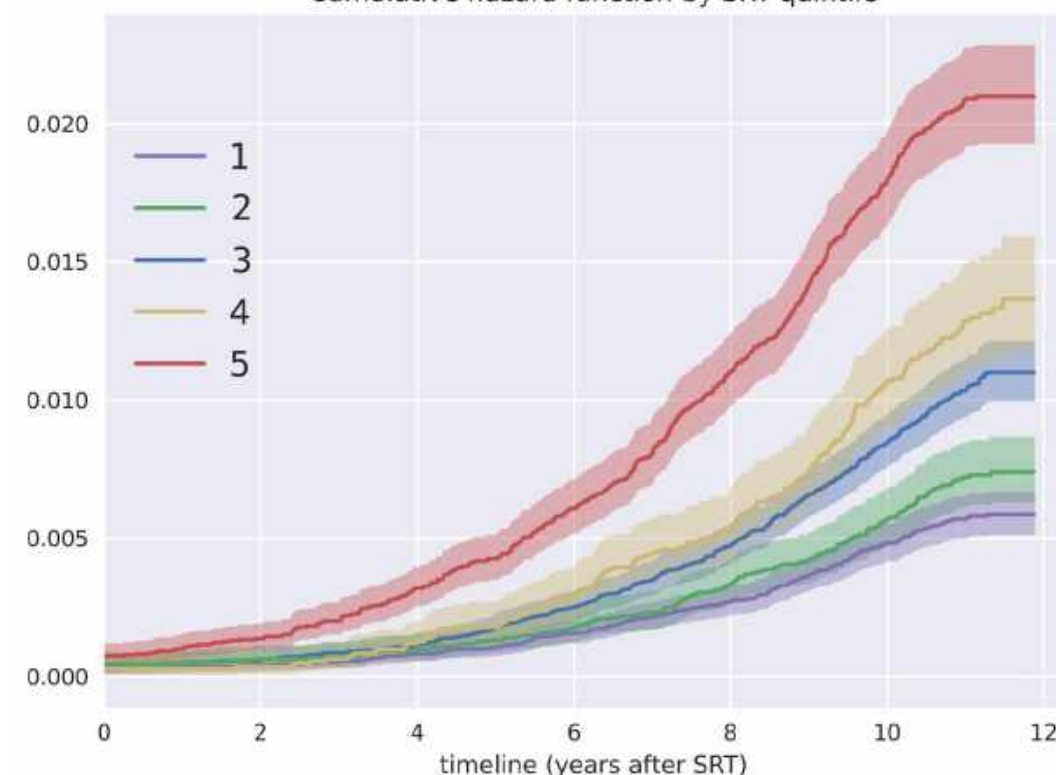
DIGITAL ASSESSMENT OF AUDITORY PERCEPTION IN DEMENTIA

NIHR awards £1.8 million to develop digital hearing tests for dementia diagnosis

The NIHR has awarded £1.8 million to fund the DIADEM (Digital assessment of auditory perception in dementia) project, led by Dr Charles Marshall of the WIPH Preventive Neurology Unit



Cumulative hazard function by SRT quintile





Article

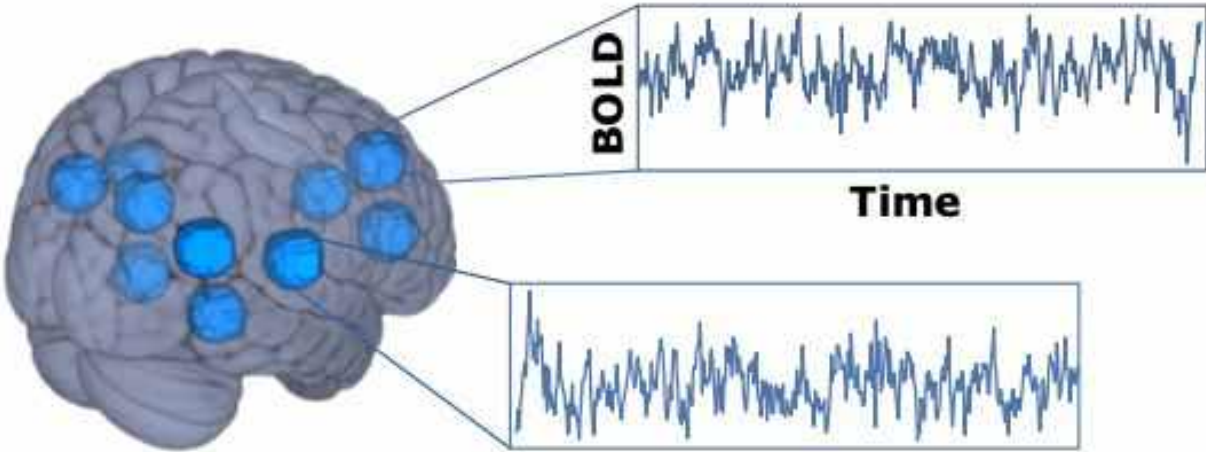
<https://doi.org/10.1038/s44220-024-00259-5>

Early detection of dementia with default-mode network effective connectivity

Received: 26 October 2023

Sam Ereira^{1,2}, Sheena Waters¹, Adeel Razi^{3,4,5} & Charles R. Marshall^{1,6}✉

Accepted: 25 April 2024



News

Opinion

Sport

Culture

Lifestyle

⋮

UK UK politics Education Media Society Law Scotland Wales Northern Ireland

The Guardian

UK

Dementia

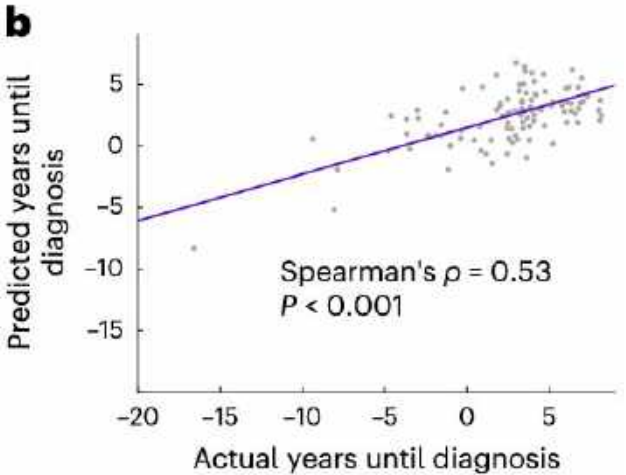
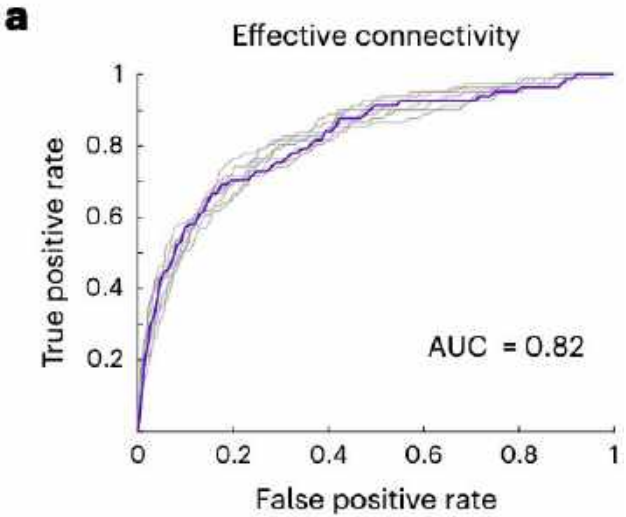
This article is more than 10 months old

Ten-minute brain scan could detect dementia early, study suggests

Scan could become routine procedure in memory clinics if findings confirmed in larger cohort, scientists say

Hannah Devlin Science correspondent

Thu 6 Jun 2024 11.58 BST



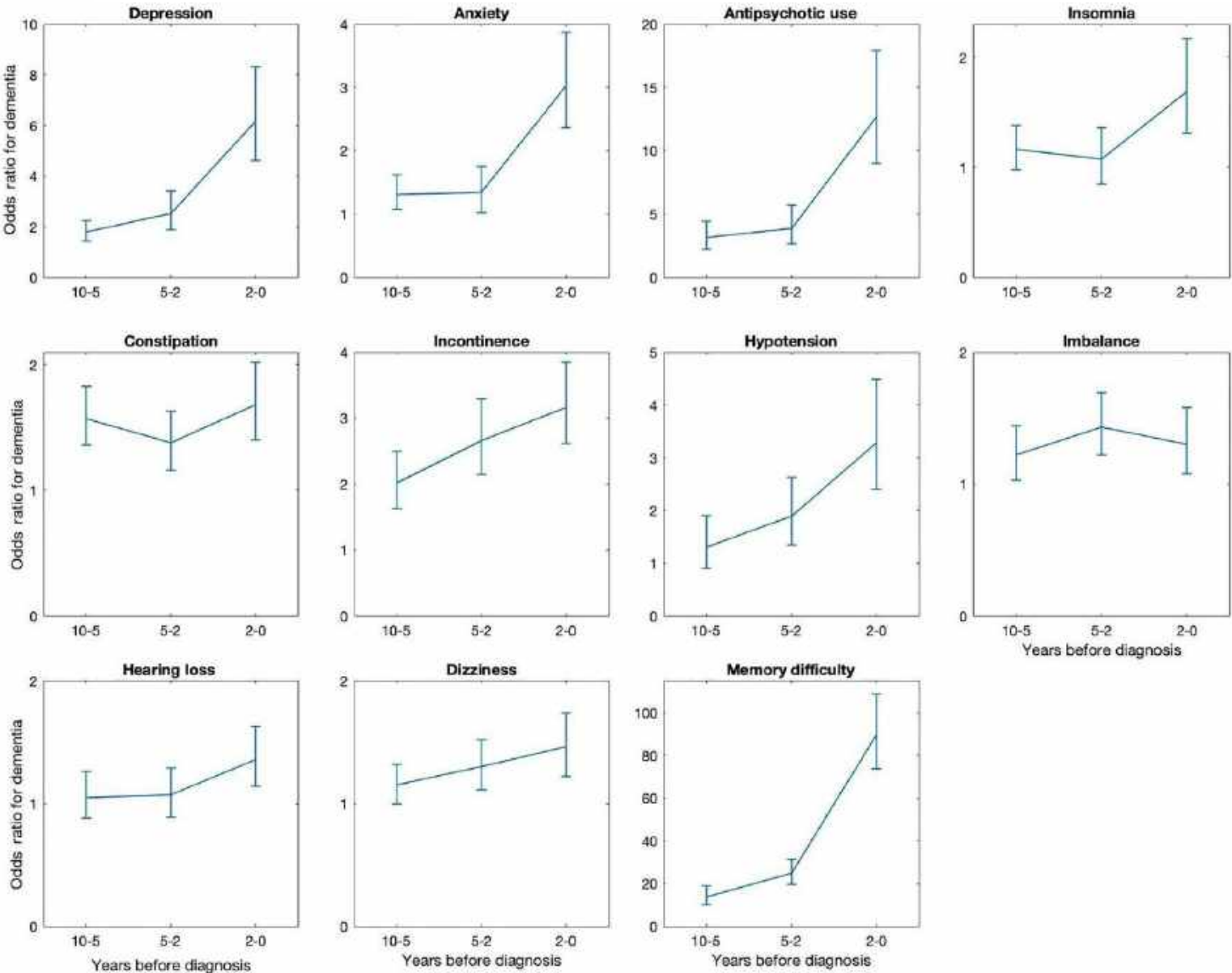
RESEARCH ARTICLE

Early presentations of dementia in a diverse population

Sedigheh Zabihi^{1,2} | Jonathan P Bestwick¹ | Mark Jitlal¹ | Phazha LK Bothongo³
Qiqi Zhang¹ | Christine Carter² | Moïse Roche² | Sarah Morgan-Trimmer⁴ |
Yvonne Birks⁵ | Mark Wilberforce⁵ | Ruth Dobson¹ | Alastair J Noyce¹ |
John Robson⁶ | Fiona M Walter⁷ | Claudia Cooper² | Charles R Marshall¹

Study reveals early dementia symptoms may vary across different ethnicities

People from different ethnic backgrounds may report physical symptoms rather than purely cognitive difficulties, leading to delays in the diagnosis of dementia.



Opportunities for search data

Errors, repetitions, altered use of language

Content reflecting difficulty with navigation, daily living etc

Content reflecting the presence of early non-cognitive symptoms

But then what?...

Session 1: Discussion Prompts

- What are the most promising health/social science research questions that can be answered using internet search data?
- What are the main barriers preventing researchers from using this data currently?
- How can internet search data uniquely contribute to improving health outcomes compared to other data sources?
- Why are researchers not using these data more extensively?

Session 1: Summary

Considerations:

Emerging Frontiers: Search data offers unique, real-time insights into health and social behavior.

Priority Use Cases: Identified key areas like mental health, early cancer detection, and understanding social dynamics.

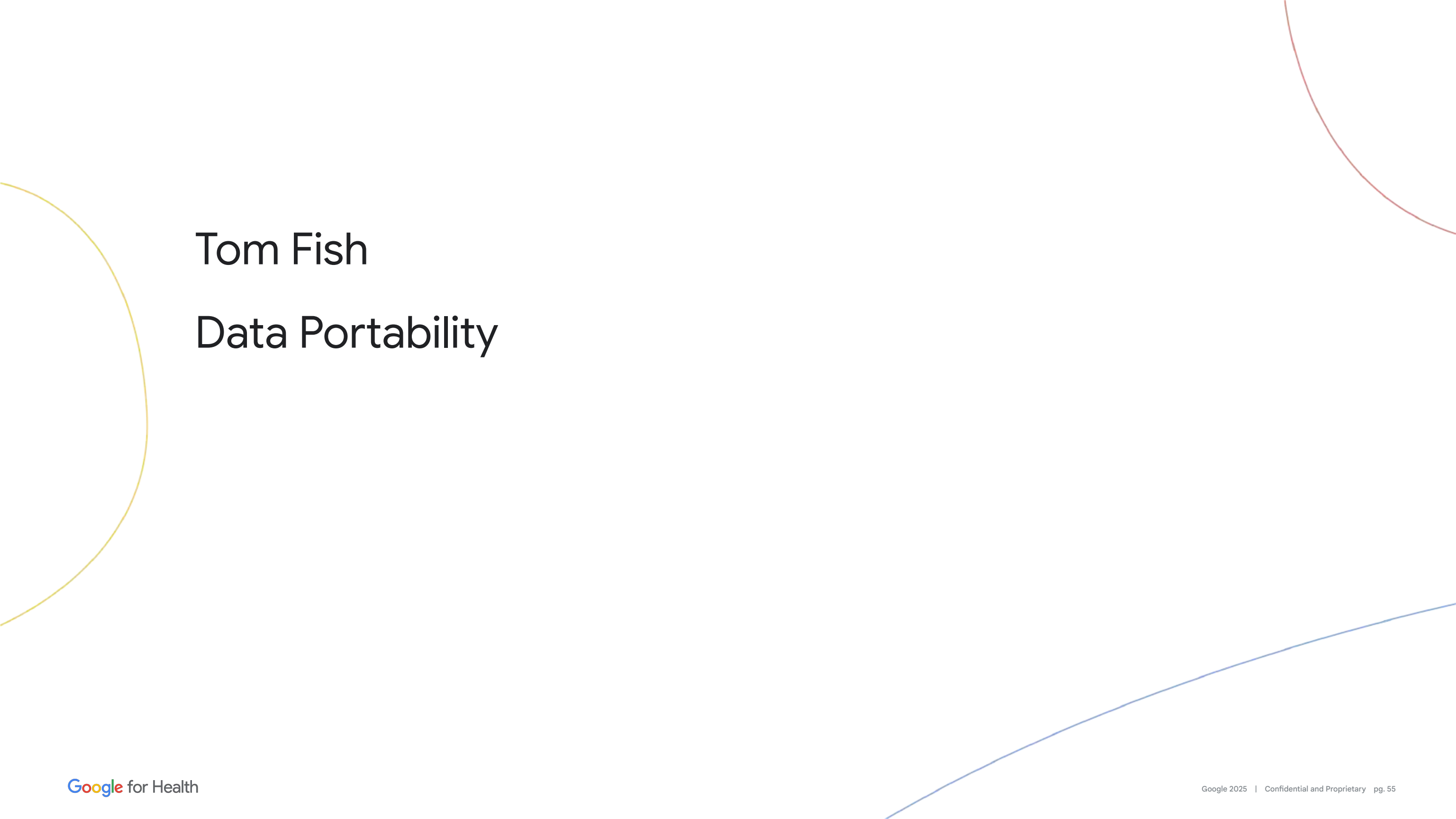
Data Portability Role: Discussed how user-controlled data donation can fuel this research.

Call for Focused Inquiry The ambition to enable "1000 new studies" by 2030

Session 2: Tools + Data Governance

Discussion Topics:

- What is available for data portability broadly in the EU and UK, across various companies.
- Google's data portability and API tools: examples of internet search data
- Ethical, legal and regulatory issues in use of internet search data



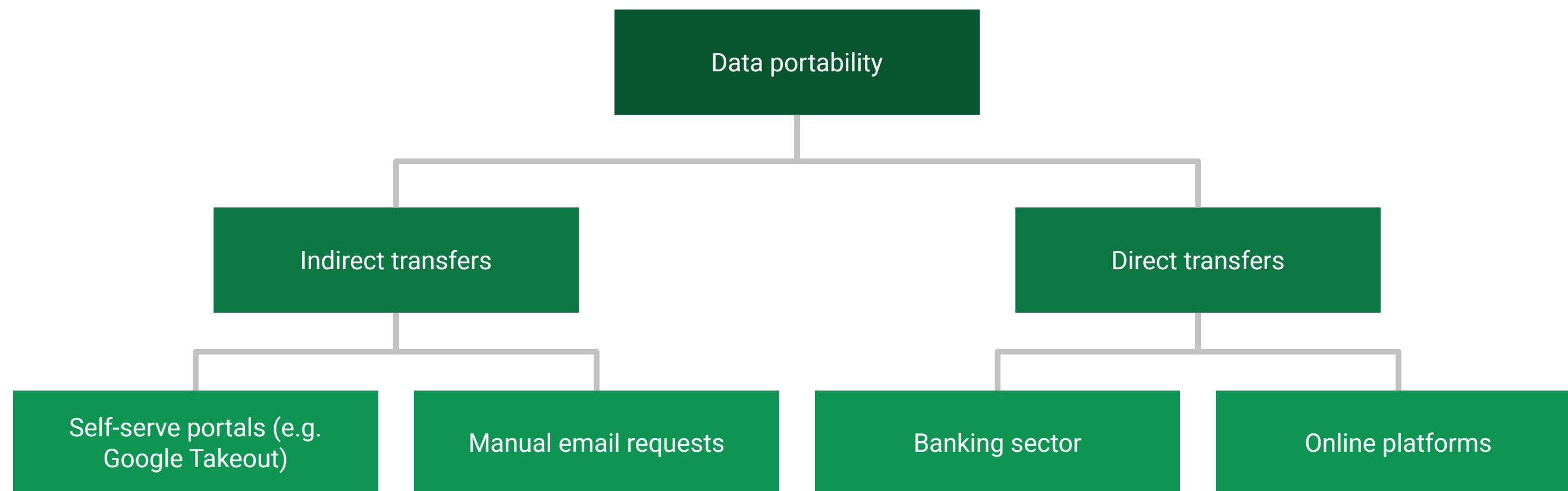
Tom Fish

Data Portability

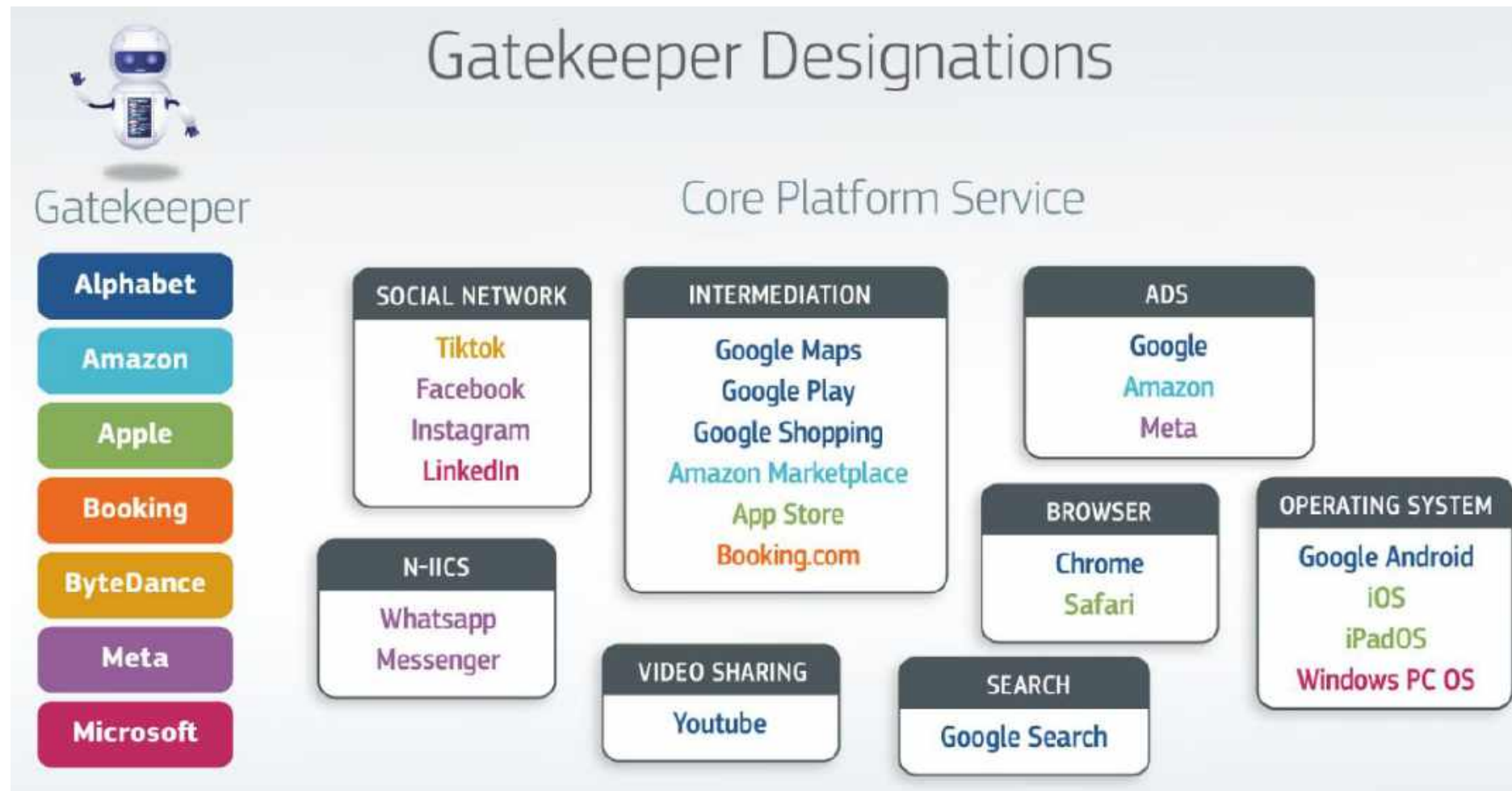


**Tools and frameworks supporting
responsible data use and portability**

What is “data portability”?



Direct transfers from online platforms: DMA



Common features:

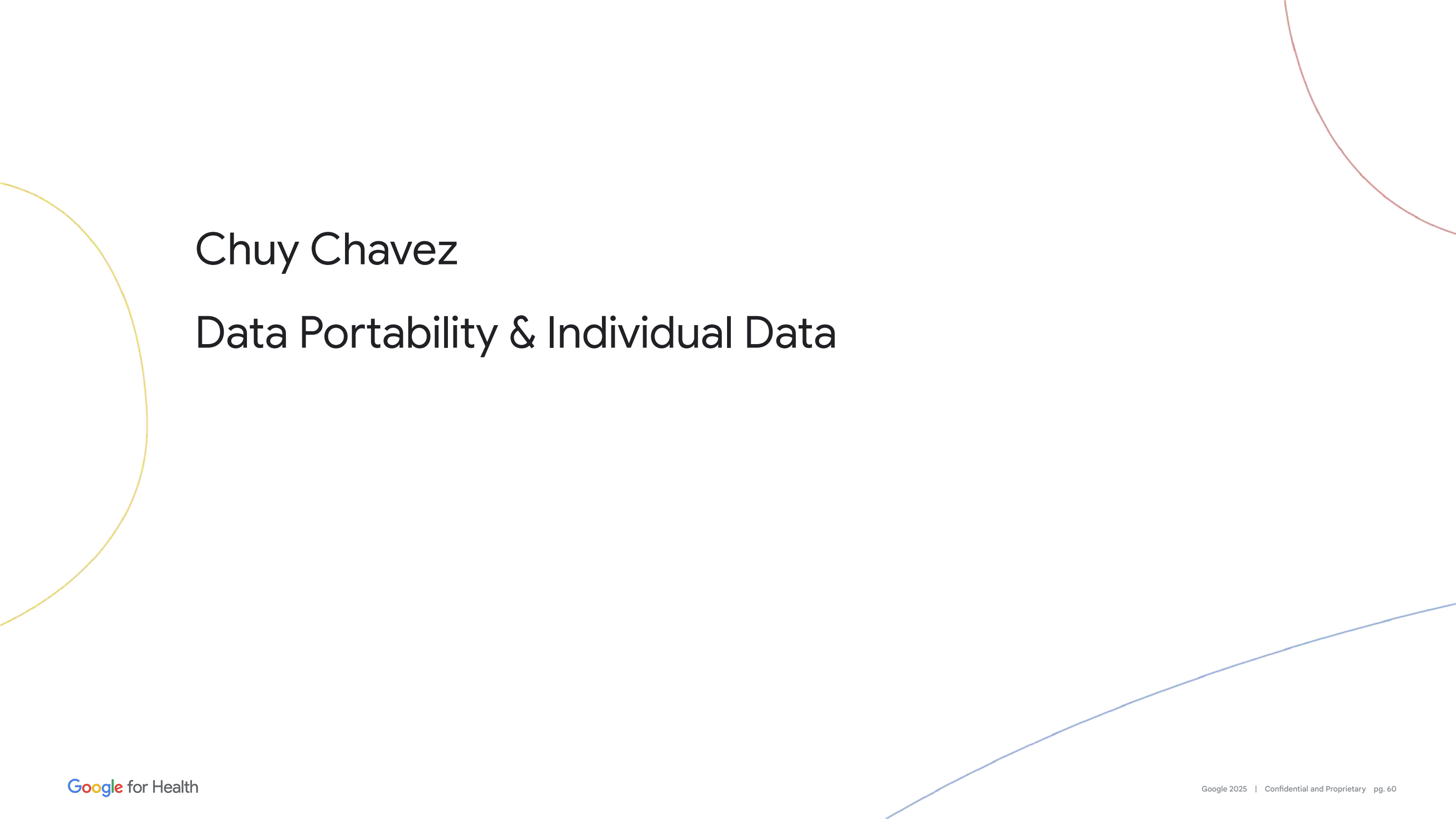
- **Frequency:** one off or ongoing daily transfers
- **Scope:** data provided by user or generated by their activity.
- **Destinations:** Third party services register separately with each gatekeeper to gain API keys.

Establishing trust for direct transfers

- Data portability reduces friction
- Guardrails are essential
- Trust managed central in the Open Banking ecosystem
- But the DMA was silent on trust

DTI's Trust Registry is in pilot phase, with support from Google.



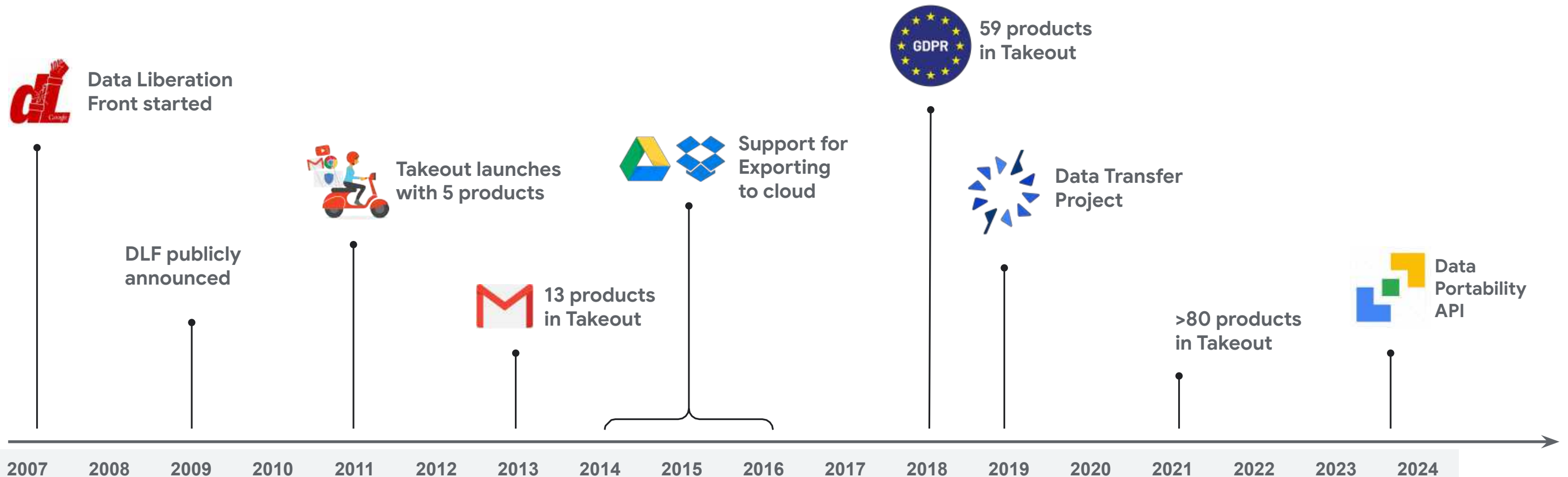


Chuy Chavez

Data Portability & Individual Data

Google's Long-standing Commitment to Data Portability

Google is a **pioneer** in data portability: for over 15 years, it has been providing **free tools** that allow for **easy user data portability**. We have developed user-facing solutions (Google Takeout), developer solutions (Data Portability API) and are a founding member of DTP enabling service-to-service portability.



Takeout gives users meaningful control of their personal data

Google Takeout:

- Is a **centralized portability tool** allowing users to export a copy of their data in commonly-used formats
- Supports **more than 80 Google product integrations** (e.g., Search, Chrome, YouTube)
- **Empowers users** to transfer data to:
 - the user's computer
 - cloud storage providers (*i.e.*, Dropbox, OneDrive, Box, and Drive)
- Includes:
 - **User-generated data** (e.g., query, visit data from Search)
 - **User-provided data** (e.g., account and profile information, ratings, reviews, saved links)



Exporting data via Takeout is intuitive, quick, and easy

- No material delay in responding to users' requests (95% of exports take < 4 hours)
- Download of 20+ years of Search activity for a frequent user in ~ 7 hours

Select data to include

Google Takeout

1 Select data to include 56 of 57 selected

Android Android device attributes, performance data, software versions, and account identifiers. [More Info](#) ☒

HTML format

Arts & Culture Favorites and galleries you've created on Google Arts & Culture. ☒

Multiple formats

Assistant Notes and Lists Notes and lists you have in Google Assistant. ☒

Multiple formats

Blogger Your Blogger blogs, including posts, pages, comments and videos, as well as settings and your Blogger profile. ☒

Multiple formats

Decide frequency and destination of the export

2 Choose file type, frequency & destination

Delivery method

Send download link via email

When your files are ready, you'll get an email with a download link. You'll have one week to download your files.

Frequency

☒ Export once
1 export

☐ Export every 2 months for 1 year
6 exports

File type & size

.zip

Zip files can be opened on almost any computer.

2 GB

Exports larger than this size will be split into multiple files.

Decide the file formats

Google Takeout

1 Select data to include

Drive options

Choose specific formats for your archive

Edit file formats

Documents **DOCX**
A document created by Microsoft Word or another word processing program.

Drawings **JPG**

Forms **XLSX**
An Excel spreadsheet created by Microsoft Excel or another spreadsheet program.

Presentations **PPTX**
A presentation file created by Microsoft PowerPoint or another presentation program.

Spreadsheets **XLSX**
An Excel spreadsheet created by Microsoft Excel or another spreadsheet program.

Uploads Original uploaded format.
Files you have uploaded to Google Drive.

Cancel OK

Drive Files you own that have been stored in your My Drive and Computers. [More info](#) ☒

Multiple formats **Advanced settings** **All Drive data included**

Customizable experiences

- The Google Takeout user interface supports parameters, so apps can customize the user interface. For example, apps can select specific products, the destination for cloud exports, and the frequency for scheduled takeouts.

The image shows a two-step process for creating a new export in Google Takeout. Step 1, 'Select data to include', shows 'My Activity' selected under the 'Products' section. Step 2, 'Choose file type, frequency & destination', shows 'Add to Drive' as the destination and 'Export every 2 months for 1 year' as the frequency.

CREATE A NEW EXPORT

1 Select data to include 1 of 1 selected

Products Deselect all

My Activity
Records of your activity data, along with image and audio attachments. [More info](#) ☒

Multiple formats **All activity data included**

[Show more products](#) **Next step**

2 Choose file type, frequency & destination

Destination

Transfer to:

Add to Drive

After we finish exporting your copy, we will add your files to [Drive](#) and email you a link to their location. These files will count against your [storage quota](#). [Learn more](#) about how to locate, access, and share your data.

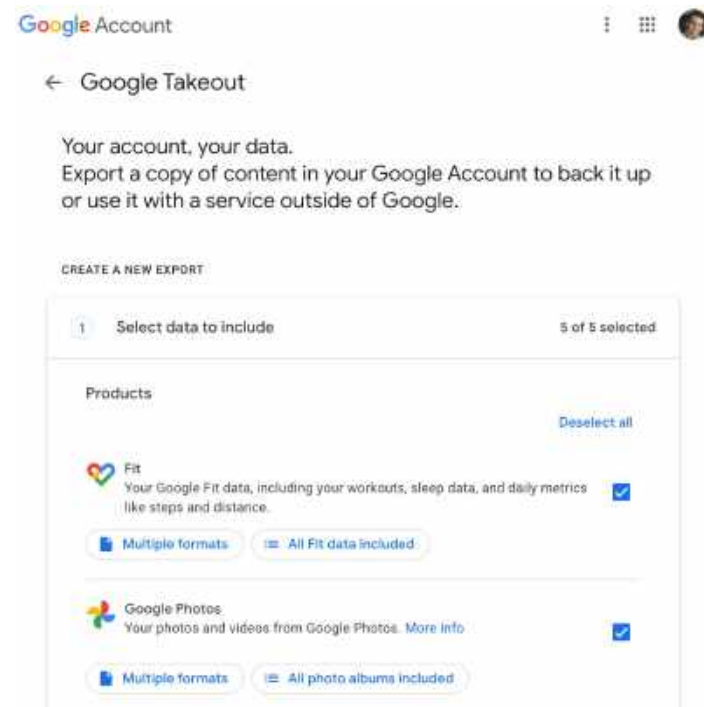
Frequency

☐ Export once
1 export

☒ **Export every 2 months for 1 year**
6 exports

Users can use Takeout to export data into a dropzone

- Once users select a cloud storage provider as the destination for their exported files, they can easily authorize a 3P to access the cloud storage and pull the data from the dropzone.

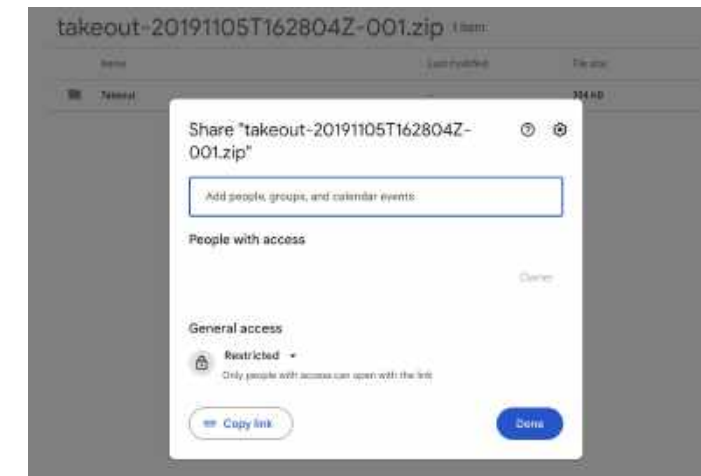


Google Takeout

Step
(1)



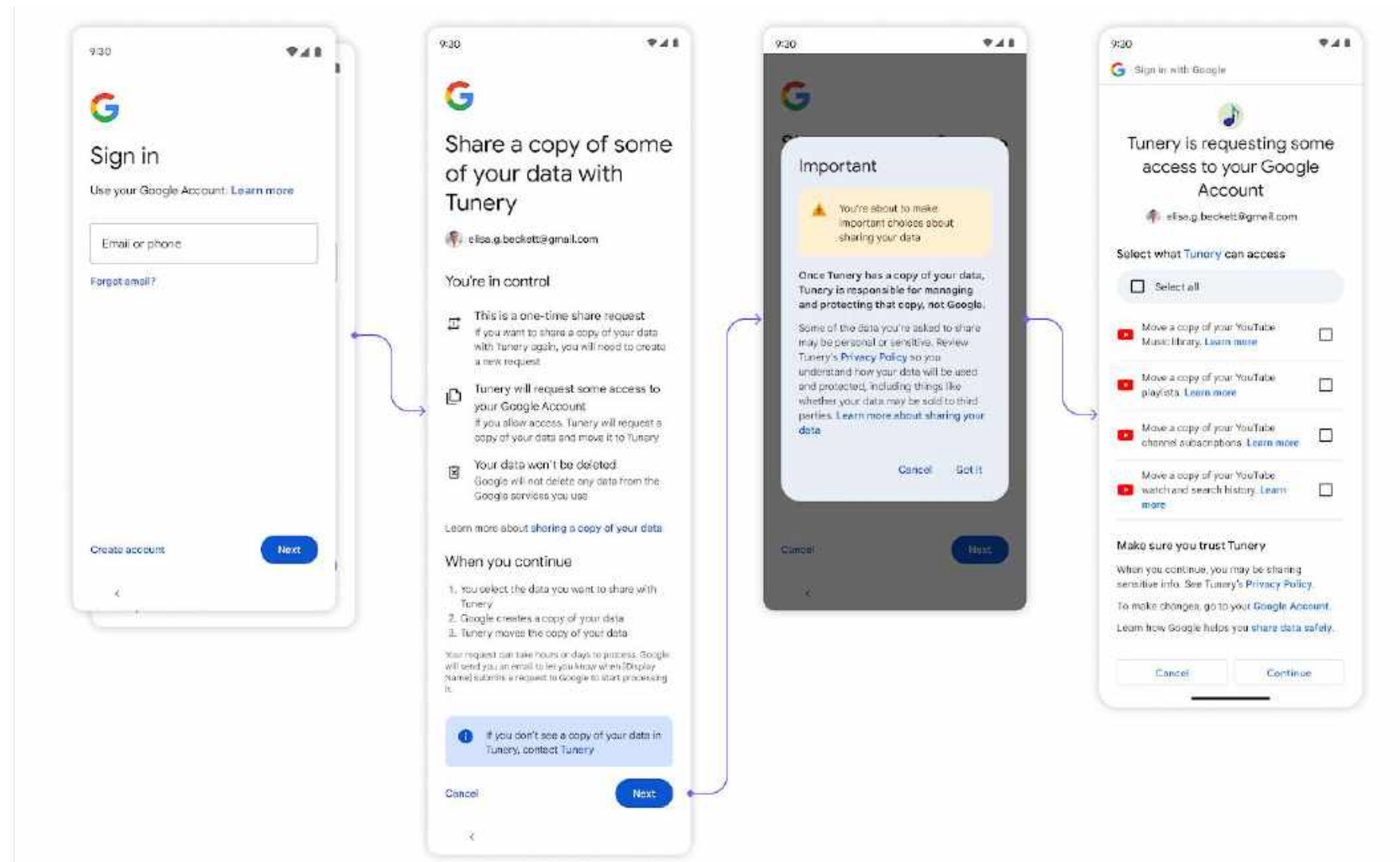
Step
(2)



Share

The Data Portability API is a powerful developer tool

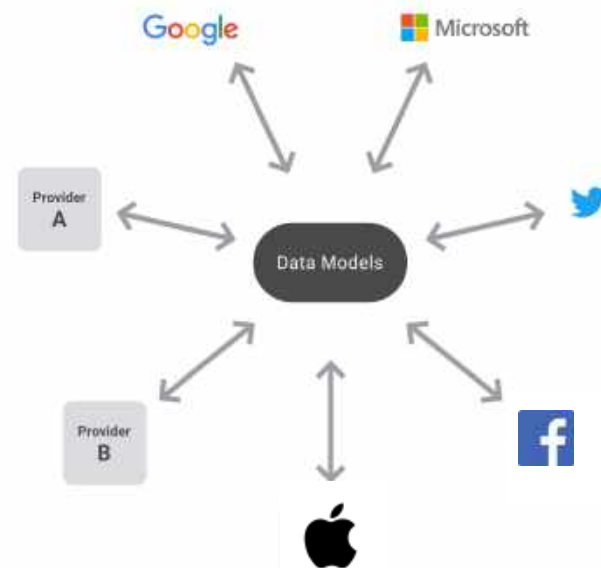
- Third Parties and developers can leverage the Data Portability API to create innovative products and services with a seamless user experience.



Data Transfer Project

DTP makes it easier to scale service-to-service portability across the internet

- open to all companies
- scalable
- promotes innovation
- reciprocal



About us

The Data Transfer Project was launched in 2018 to create an open-source, service-to-service data portability platform so that all individuals across the web could easily move their data between online service providers whenever they want.

The contributors to the Data Transfer Project believe portability and interoperability are central to innovation. Making it easier for individuals to choose among services facilitates competition, empowers individuals to try new services and enables them to choose the offering that best suits their needs.

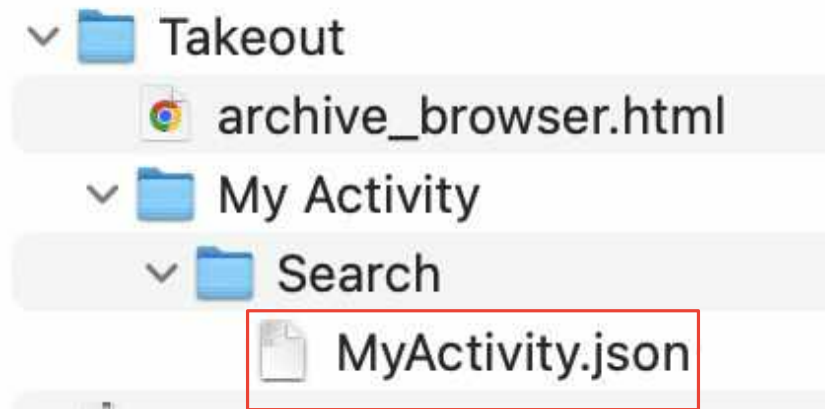
Current contributors include:



Google Takeout & API Comparison

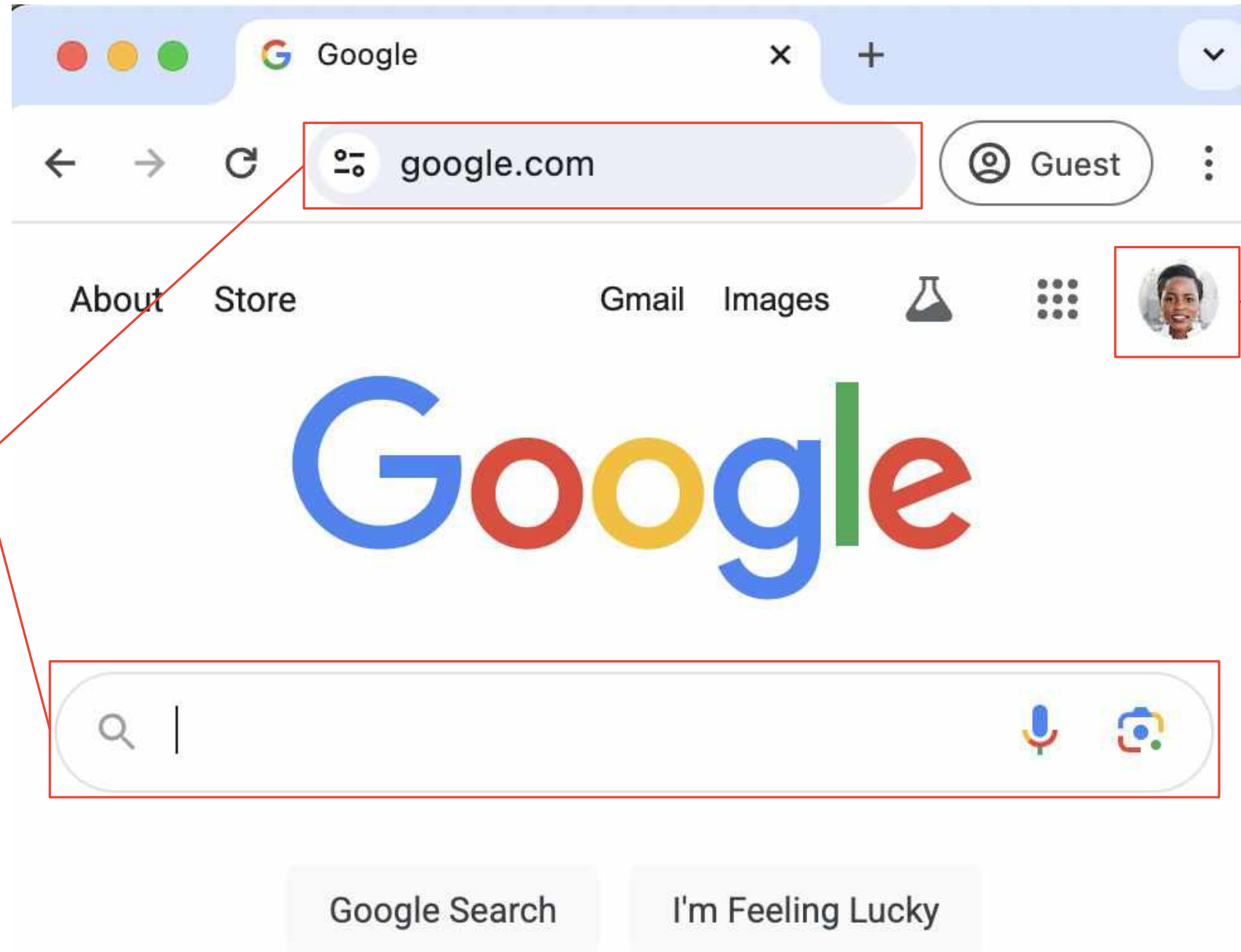
Google Takeout	Data Portability API
End user product	Developer product
Available to end users Globally	Available to end users in EU, UK, and Switzerland
Data available from 80+ product areas including search activity	Data available from 8 product areas including search activity
Data is provided in machine or human readable formats	Data is provided in machine readable formats
Requires data analysis skills to parse json files	Requires data analysis skills to parse json files and software/app development skills to use API

Example Search data



```
[{
  "header": "Search",
  "title": "Visited Low Back Pain Pictures: Symptoms, Causes, Treatments - WebMD",
  "titleUrl":
    "https://www.webmd.com/back-pain/ss/slideshow-low-back-pain-overview",
  "time": "2024-04-30T19:47:30.267Z",
  "products": ["Search"],
  "activityControls": ["Web \u0026 App Activity"]
},{
  "header": "Search",
  "title": "Searched for lower back pain",
  "titleUrl": "https://www.google.com/search?q\u003dlower+back+pain",
  "time": "2024-04-30T19:46:27.300Z",
  "products": ["Search"],
  "activityControls": ["Web \u0026 App Activity"]
}]
```

When Search data is available in Google Takeout



Searches in omnibox and google.com are capture in Google Takeout

Only searches when logged in are available in Google Takeout

When Search data is available in Google Takeout

1. **Google Account**

Search Google Account

Home

Personal info

2. **Data & privacy**

Security

People & sharing

Payments & subscriptions

About

Things you've done and places you've been

Your options for history, ads, and personalization. Rediscover the things you've searched for, read, and watched, and see the places you've visited.

History settings

Choose whether to save the things you do and places you go to get more relevant results, personalized maps, recommendations, and more. Location info is saved and used based on your settings. [Learn more](#)

3. **Web & App Activity** ☒ On >

Timeline ☐ Paused >

YouTube History ☒ On >

See and delete your history anytime

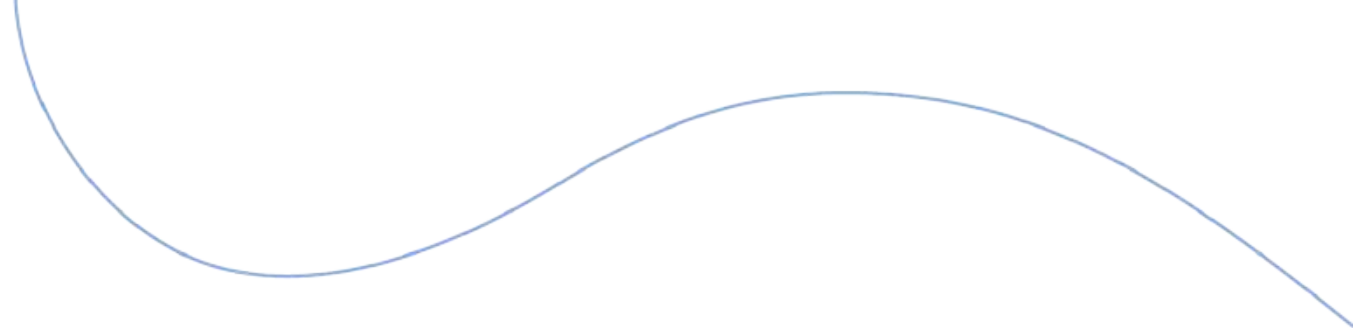
[My Activity](#) [Maps Timeline](#) [YouTube watch & search history](#)

When Search data is available in Google Takeout

Data available in Google Takeout is **independent** of:

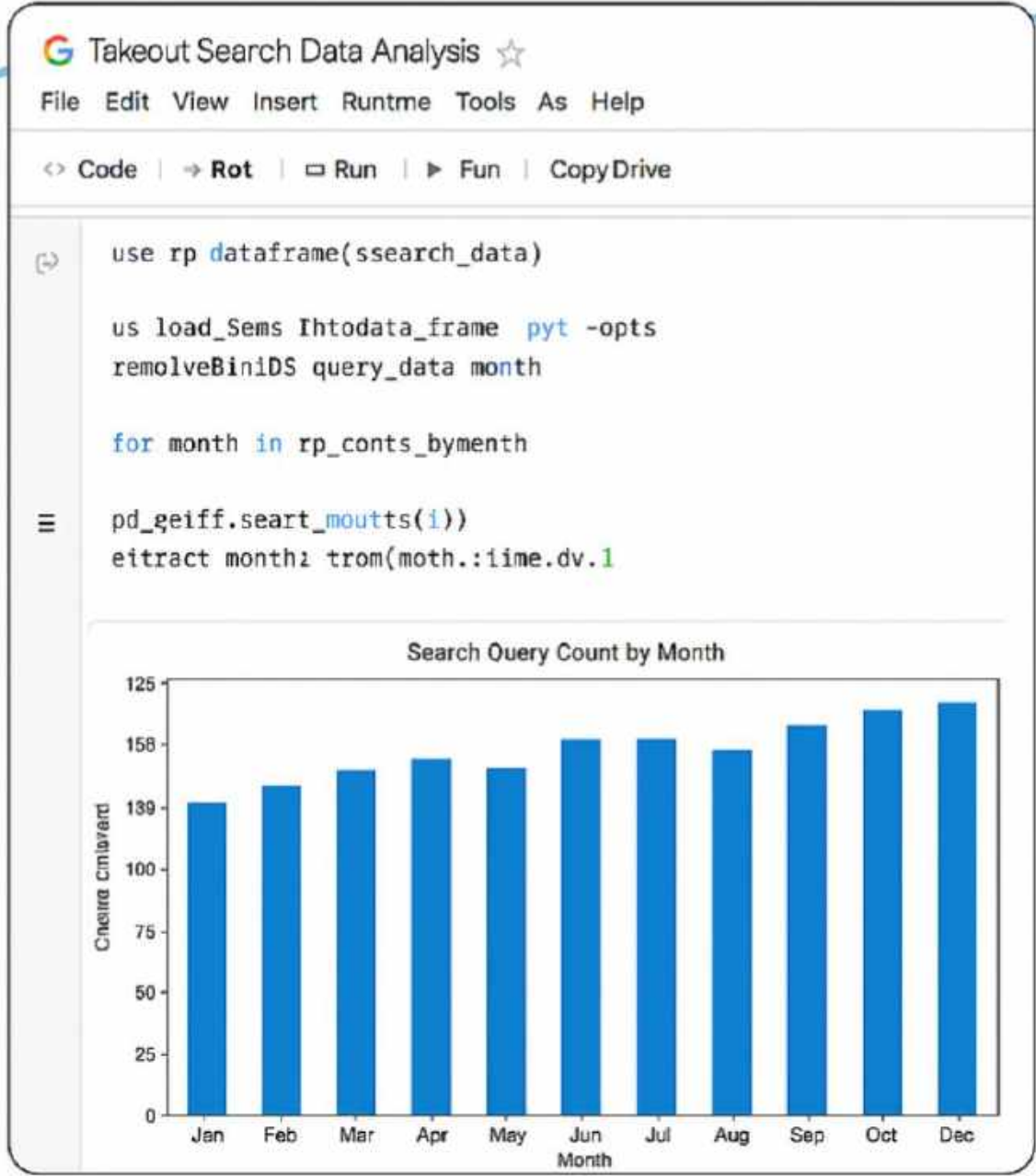
- Browser (i.e. Chrome, Safari)
- Platform (i.e. Mobile vs Desktop)
- Operating System (i.e. Android, iOS)

However, some browsers use Google Search by default in the omnibox.



Search Trends and Behavioral Metadata might indicate emerging health risks before clinical diagnosis

COLAB: Takeout Search Data



COLAB: Takeout Search Data

JSON Loading & Monthly Query Volume

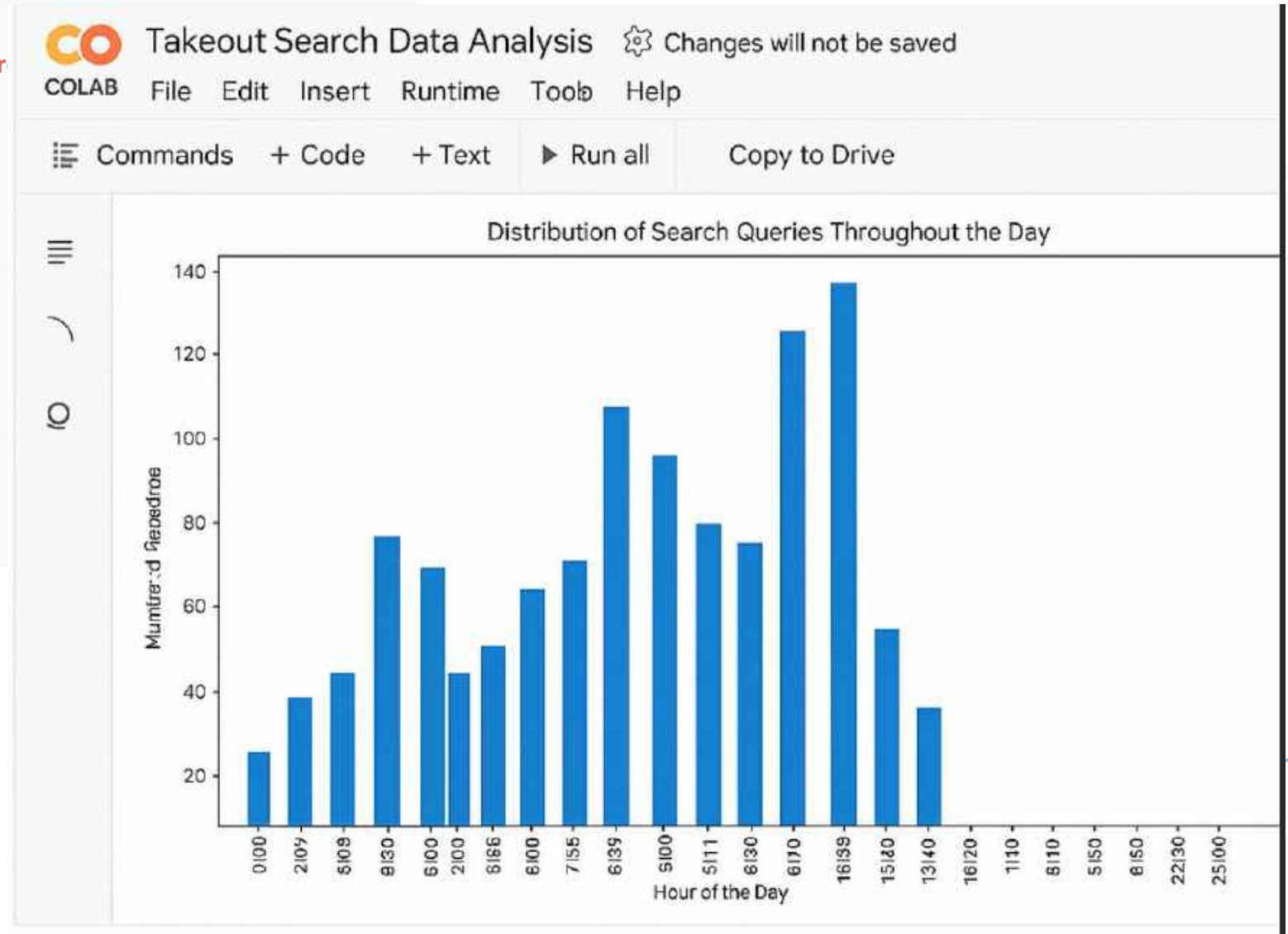
```
# prompt: using the same dataframe create a distribution graph that shows when the user usually searches for things in a given day
```

```
# Convert the 'time' column to datetime objects if not already done
df['time'] = pd.to_datetime(df['time'], format='%Y-%m-%d%H:%M:%S.%f', errors='coerce')
```

```
# Extract the hour of the day
df['hour'] = df['time'].dt.hour
```

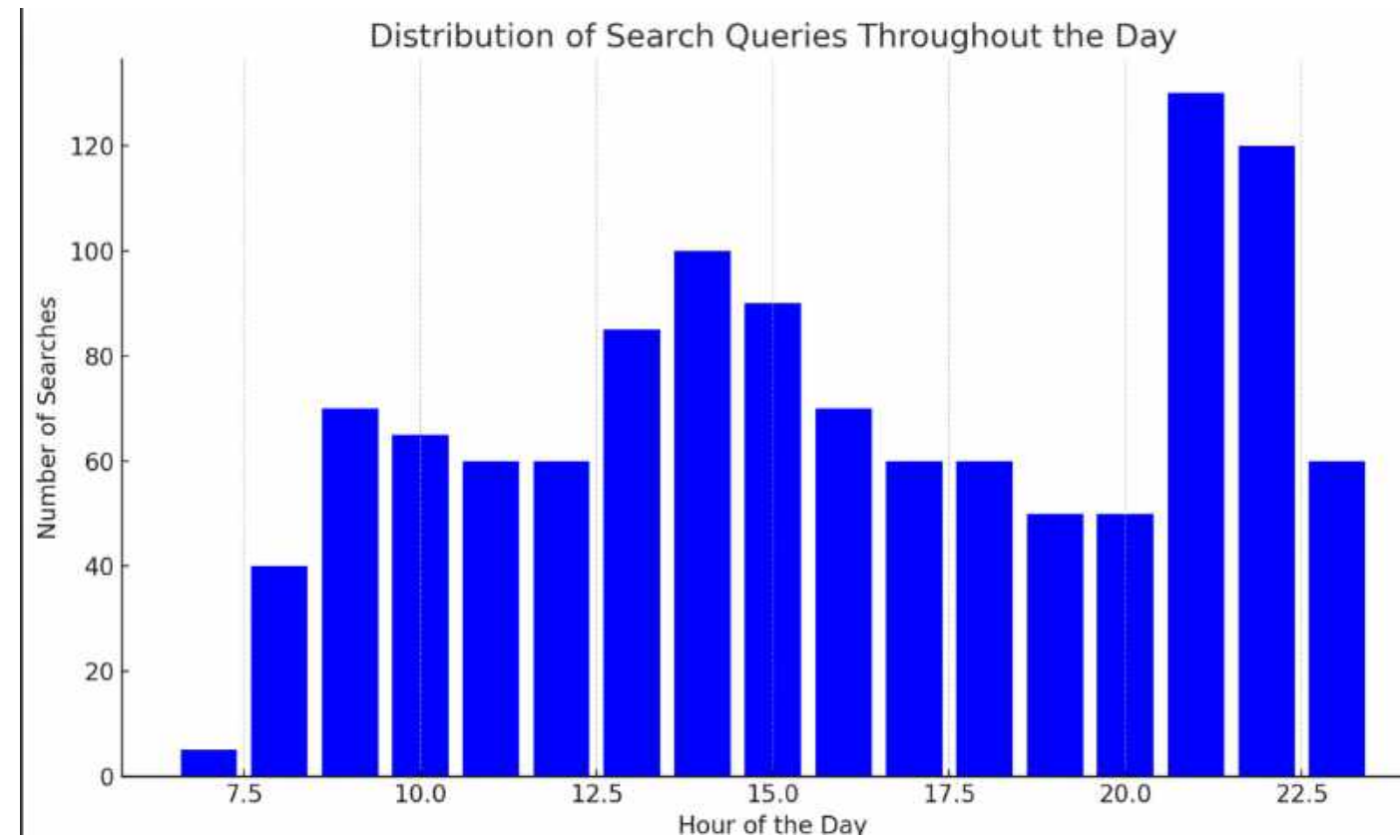
```
# Group by hour and count the searches
hourly_counts = df.groupby('hour')['time'].count()
```

```
# Create the distribution plot
plt.figure(figsize=(10, 6))
hourly_counts.plot(kind='bar')
plt.xlabel('Hour of the Day')
plt.ylabel('Number of Searches')
plt.title('Distribution of Search Queries Throughout the Day')
plt.xticks(range(24)) # Ensure all 24 hours are shown on the x-axis
plt.tight_layout()
plt.show()
```



COLAB: Takeout Search Data

Hourly Query Volume



[] # prompt: Using the same data frame count how many entries are for searches versus visits. This can be done by looking at the "title" and parsing out th

Assuming 'df' is the DataFrame from the previous code

```
def categorize_search(title):  
    """ "Categorizes a search entry as 'Search' or 'Visit' based on the title." ""  
    try:  
        first_word = title.split()[0]
```

Legal and ethical considerations in the use of internet search data.

12th June 2025

Dr Jessica Bell

Assistant Professor, University of Warwick
PI, Born in Scotland Data Trust
jessica.bell@warwick.ac.uk

Session 2: Discussion Prompts

- What are the current limitations and capabilities of data portability options in the EU and UK?
- How can Google's data portability and API tools be more effectively used by researchers?
- What are the key ethical and legal concerns associated with using individual internet search data for research?
- What governance frameworks need to be in place to ensure responsible use of donated personal data?

Session 2: Summary

Considerations:

Empowering Users: Google Takeout puts data control directly in users' hands.

API Potential: Exploring Google's API tools to facilitate secure research access.

Ethical Bedrock: Privacy, informed consent, and data security are non-negotiable.

Governance Frameworks: The need for robust guidelines for donated personal data.

Session 3: Data Integration @ Scale



**Smart
Data
Donation
Service**

**Community
infrastructure** for
scaling data access

Dr. David Zendle



Smart Data
Donation Service

David
Zendle

Core Role

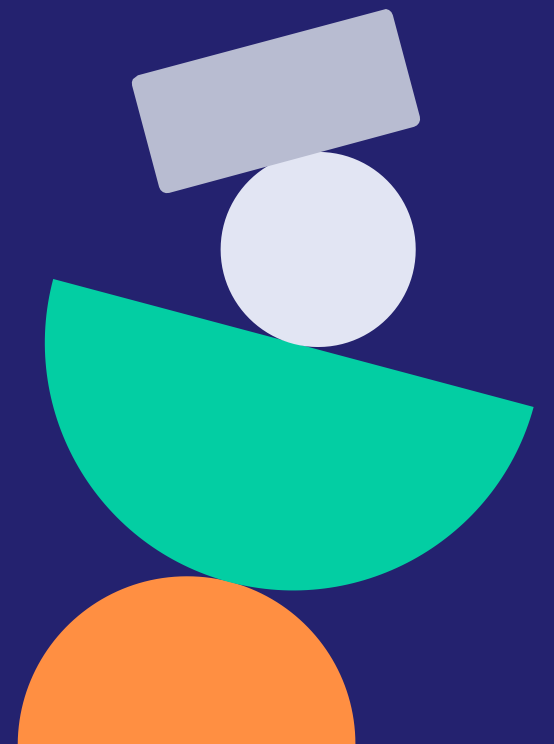
Director, Smart Data Donation Service

Advisory Roles

Advisory Board for Safer Gambling

DCMS College of Experts

Ofcom Research WG









SDDS at Scale



Smart Data
Donation Service

£7.9m
5 years

Initial Domains

Social Media & Gaming

Core Remit

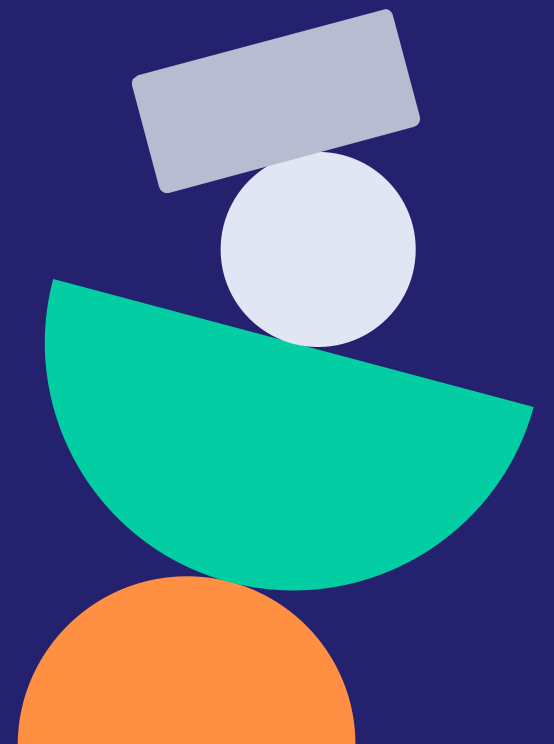
Evidence-based policy

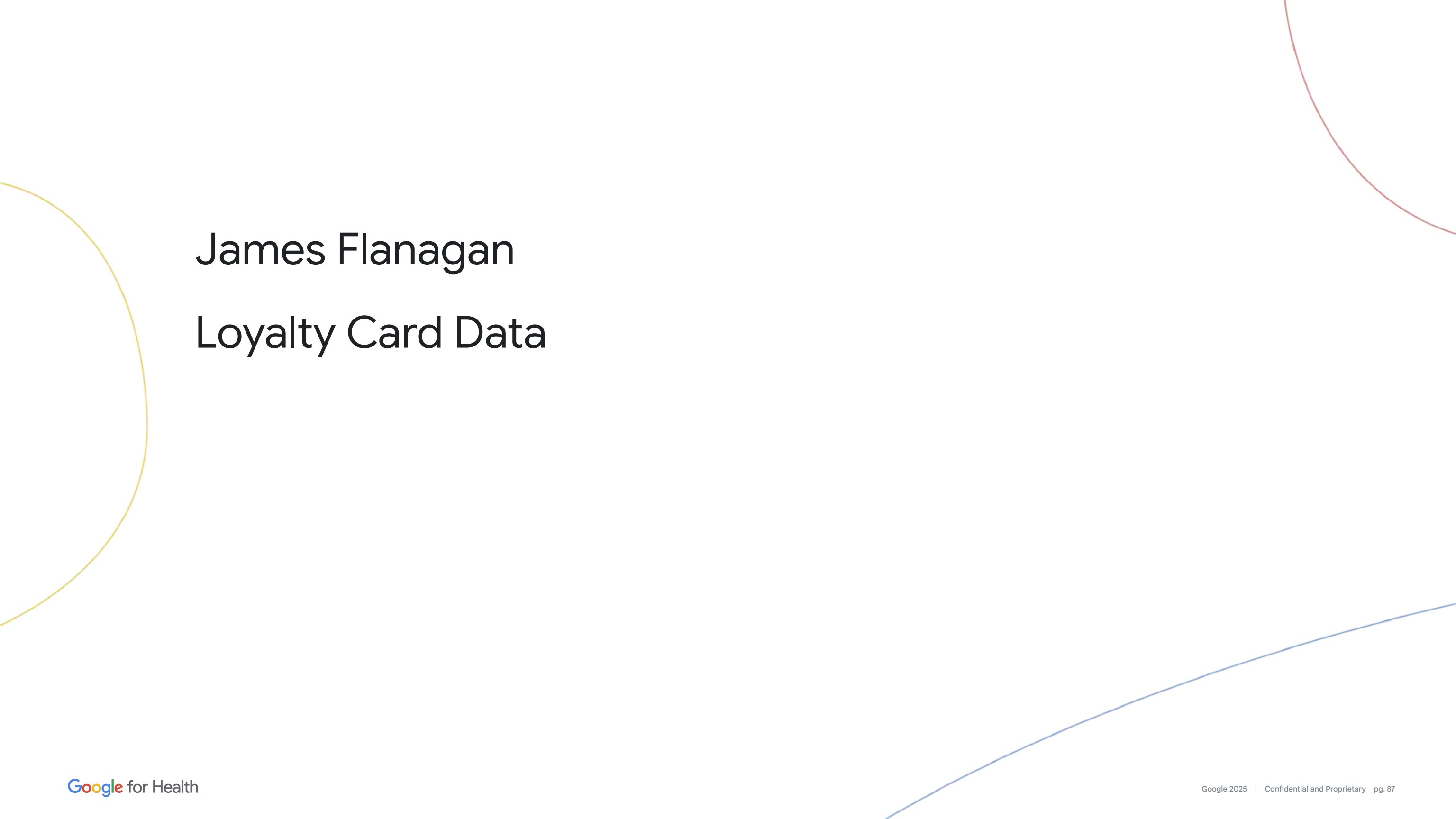
Key Scaling Mechanisms

Dedicated codesign unit

Cautious risk posture

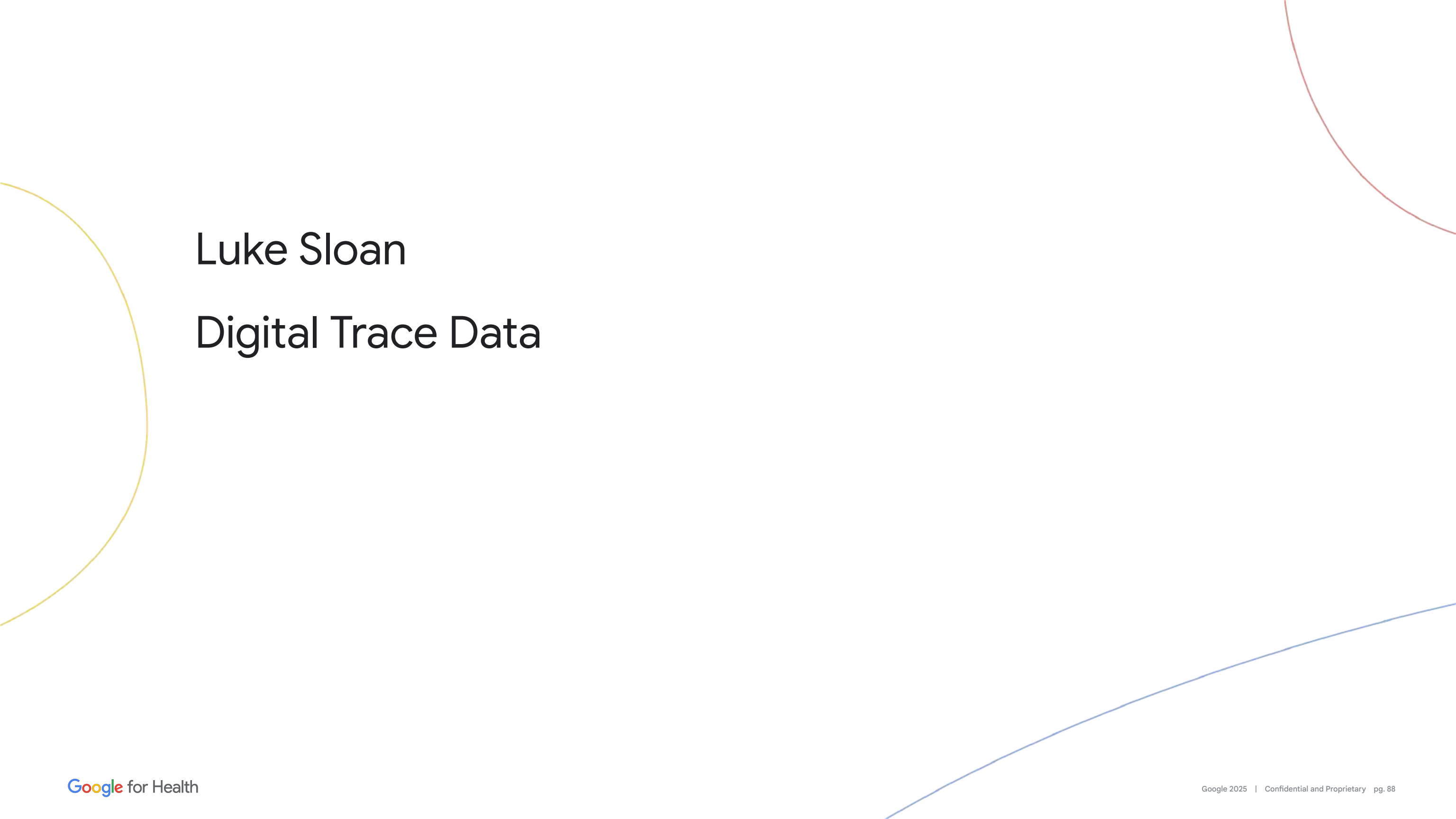
Ecosystem enriching position





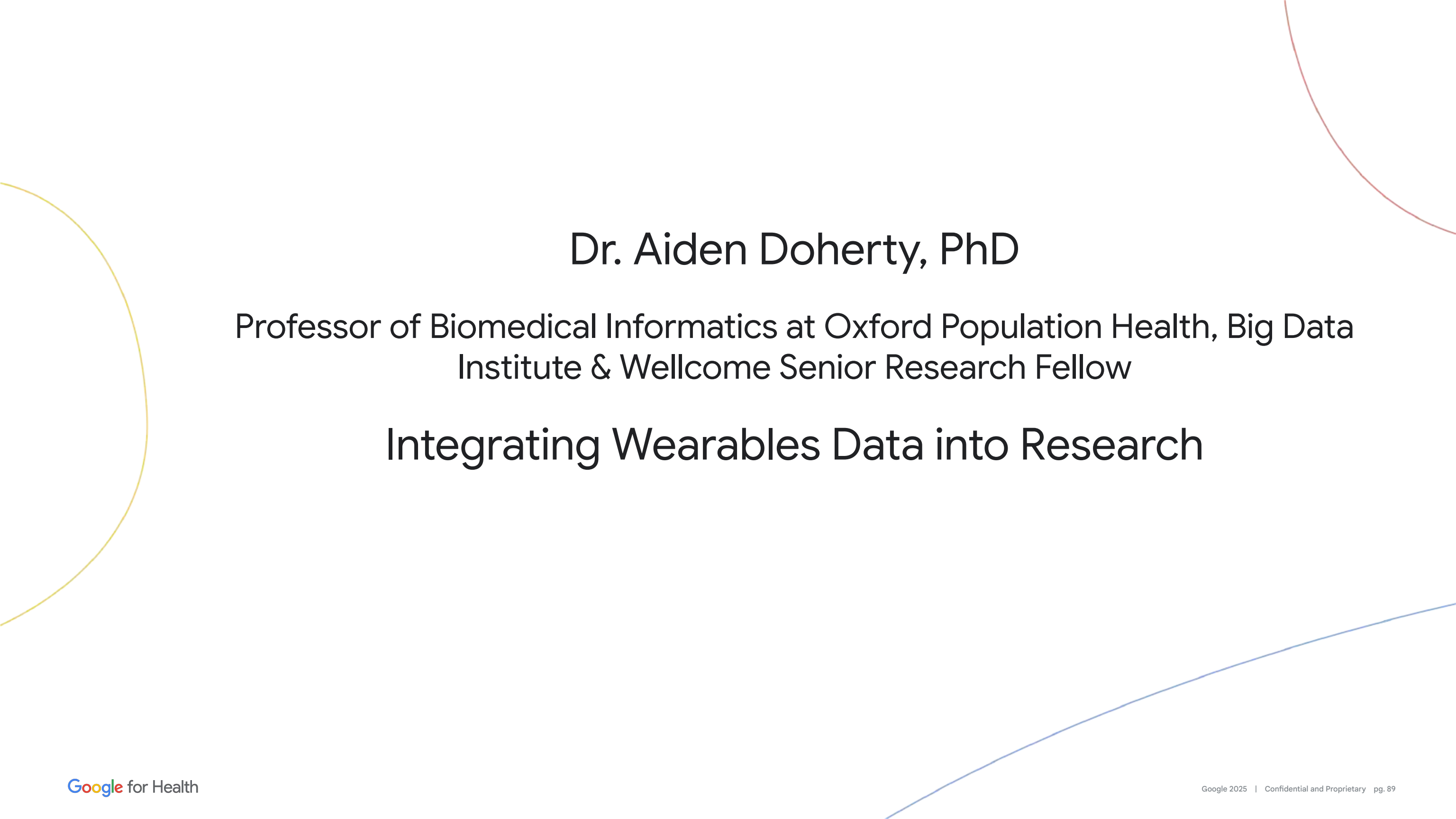
James Flanagan

Loyalty Card Data



Luke Sloan

Digital Trace Data

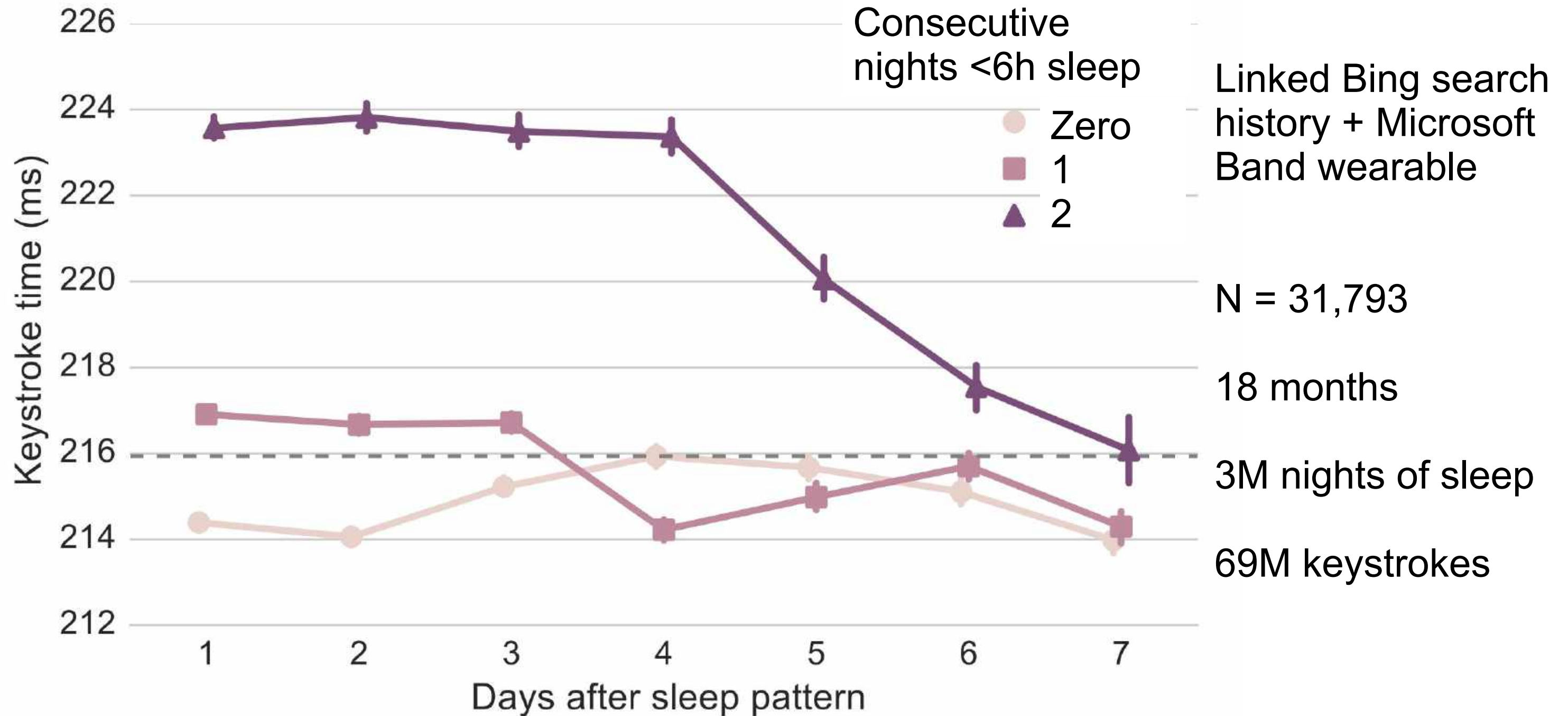


Dr. Aiden Doherty, PhD

Professor of Biomedical Informatics at Oxford Population Health, Big Data
Institute & Wellcome Senior Research Fellow

Integrating Wearables Data into Research

Search history & wearables data can be helpful for discovery research



7 day wrist-worn accelerometer data collection in UK & China

UKB 2013 – 2015
n = 103,712

Consent = 47%
Adherence = 93%

CKB 2020 - 2021
n = 20,375

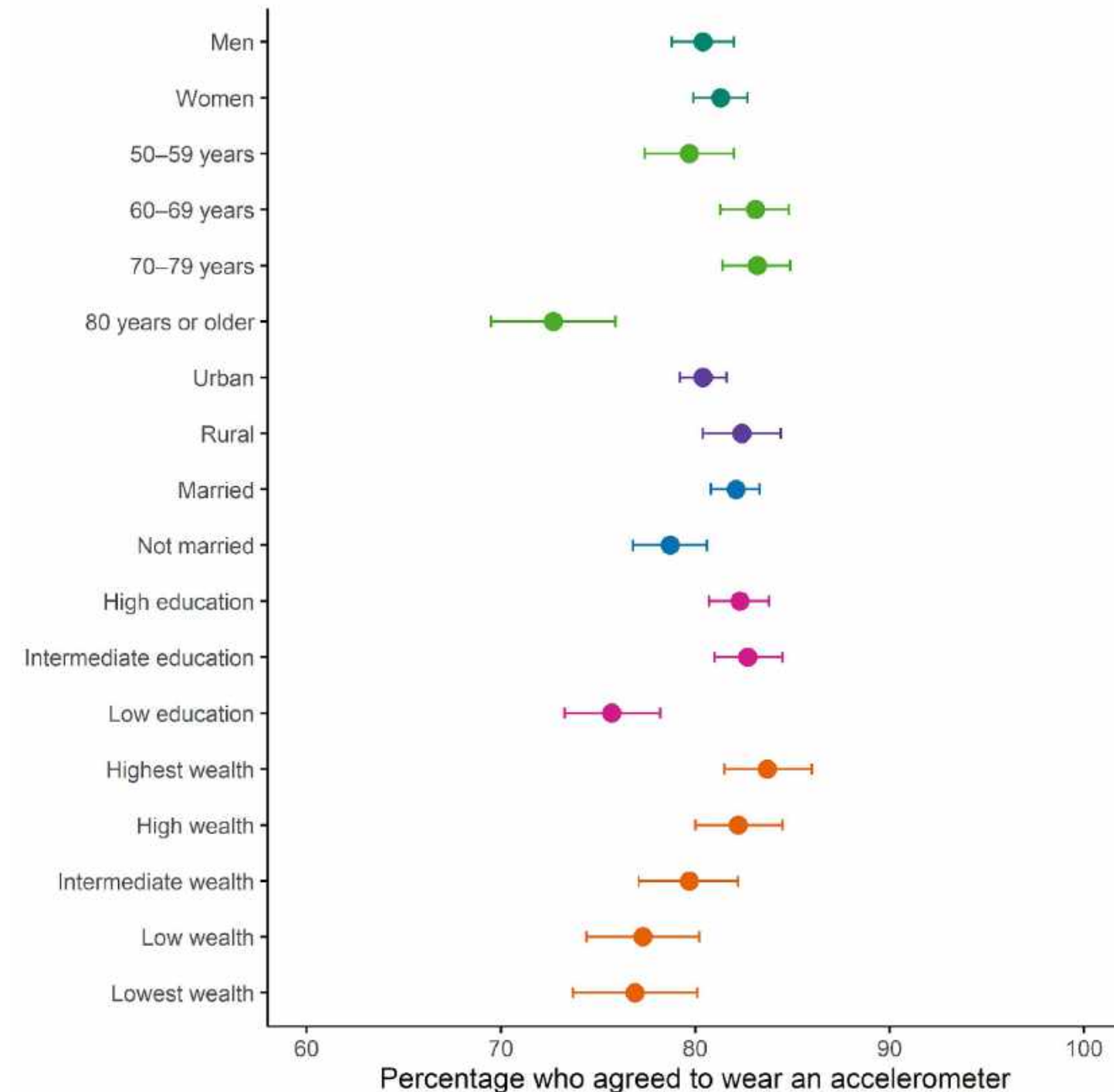
Consent = 89%
Adherence = 93%

ELSA 2021 - 2023
n = 4,354

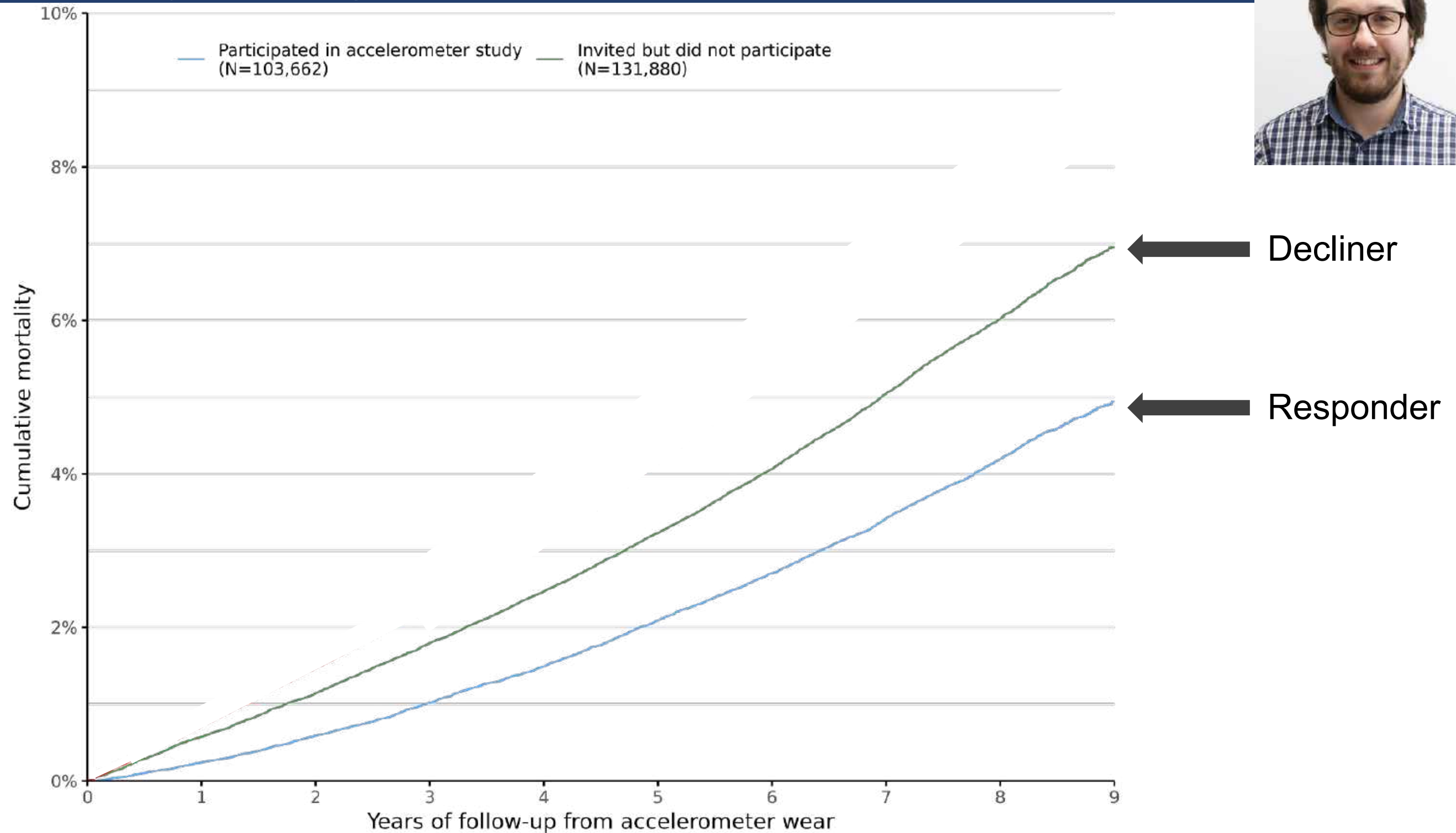
Consent = 81%
Adherence = 90%

CHARLS 2020 - 2021
n = 12,496

Consent = 76%
Adherence = 94%

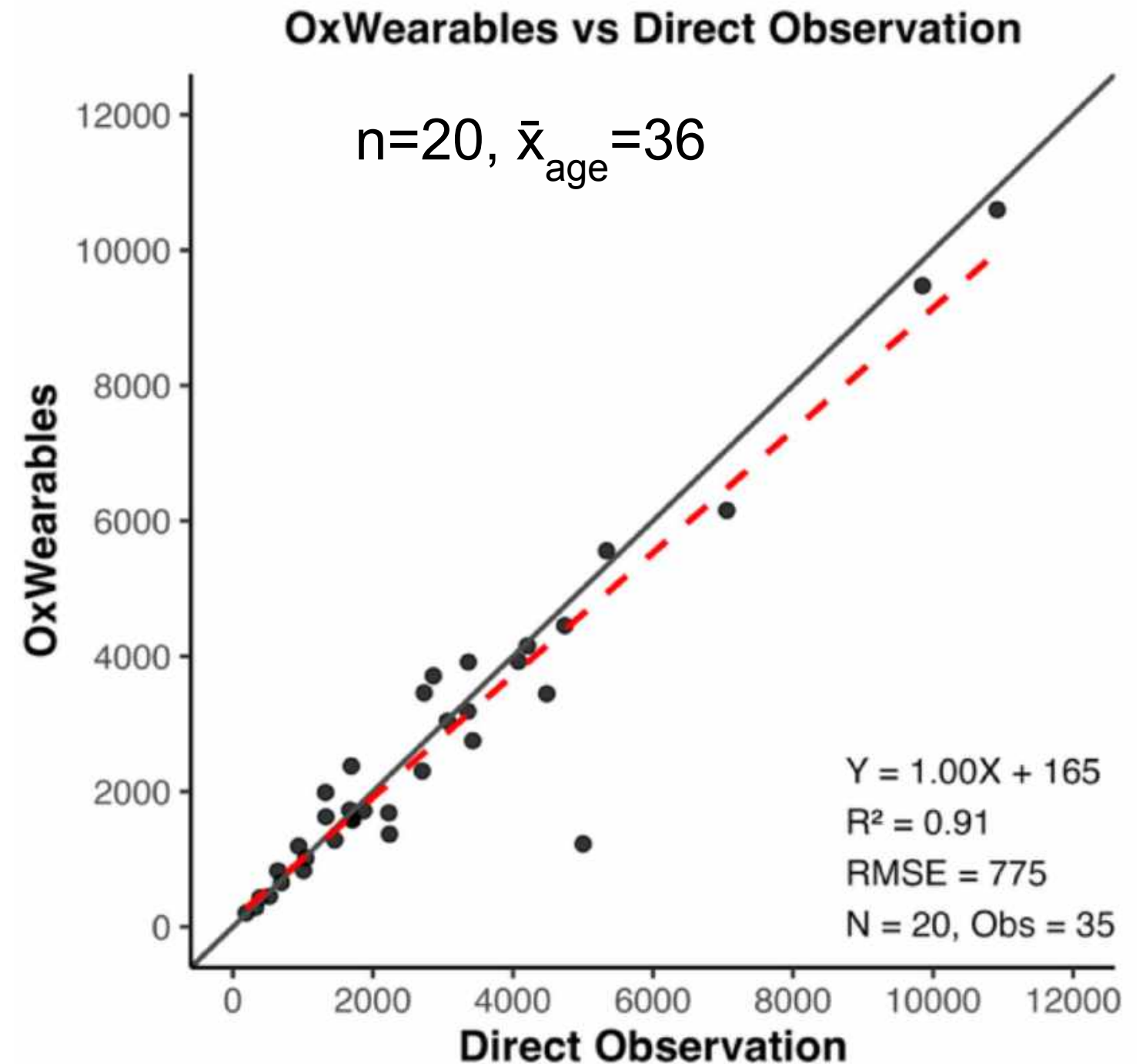
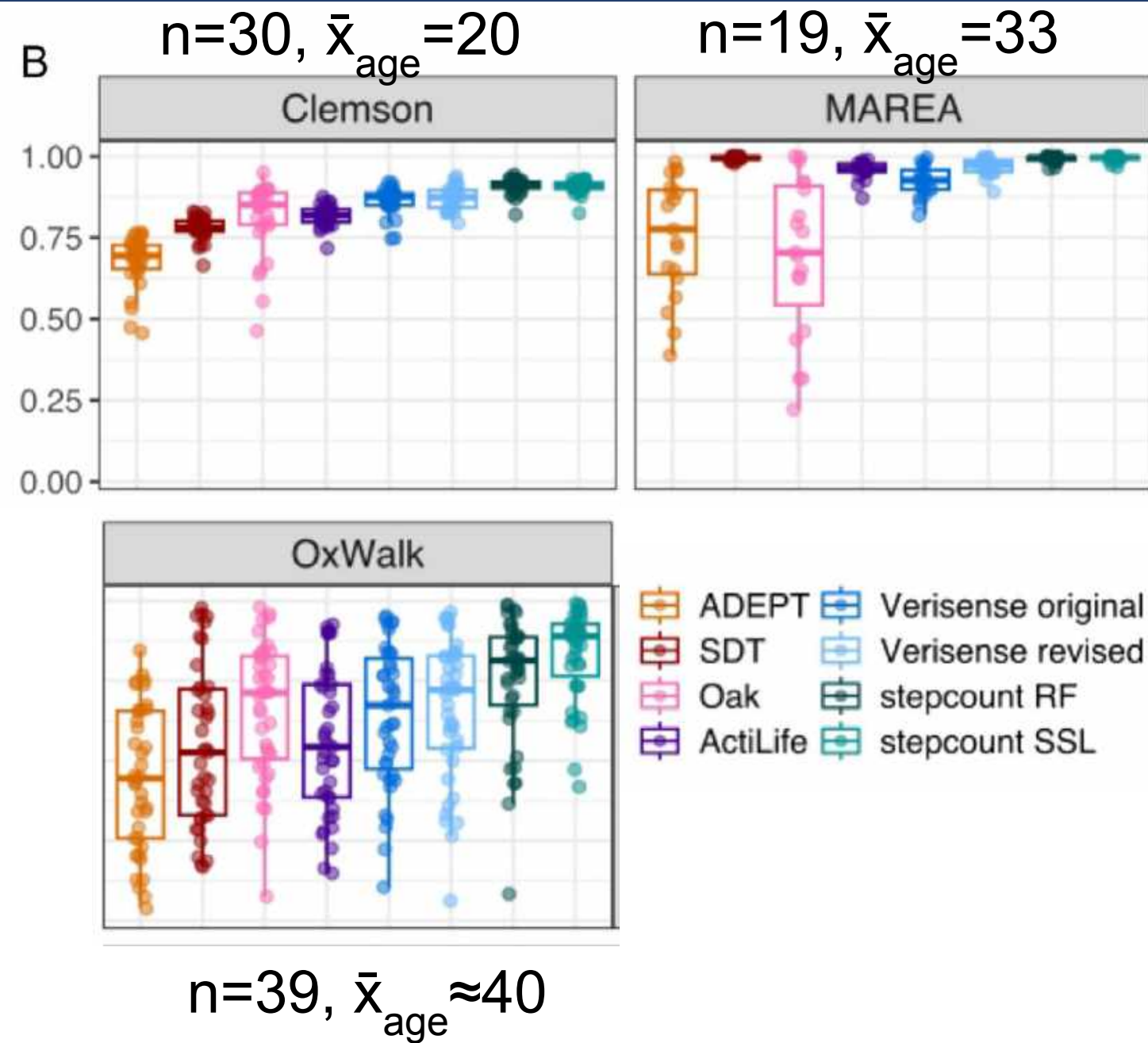


Survivorship by who agreed to wear a device in UK Biobank

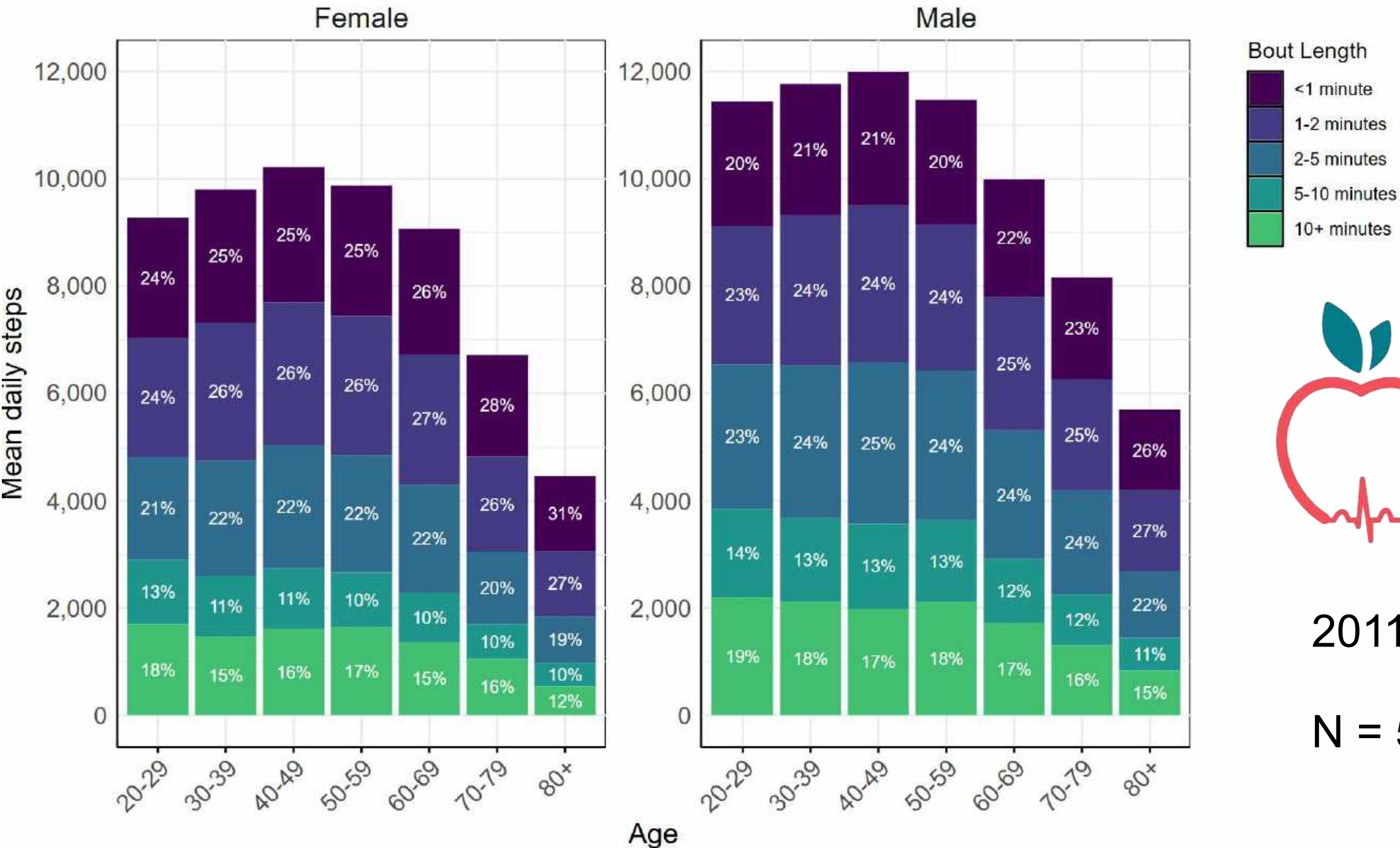


**For participants not in the accelerometer study cohort, a hypothetical study date was generated (based on the distribution of existing dates) using bootstrap sampling with replacement.*

Independent validation studies of step count measurement



Face validity in population representative samples



2011 – 2014

N = 5,153

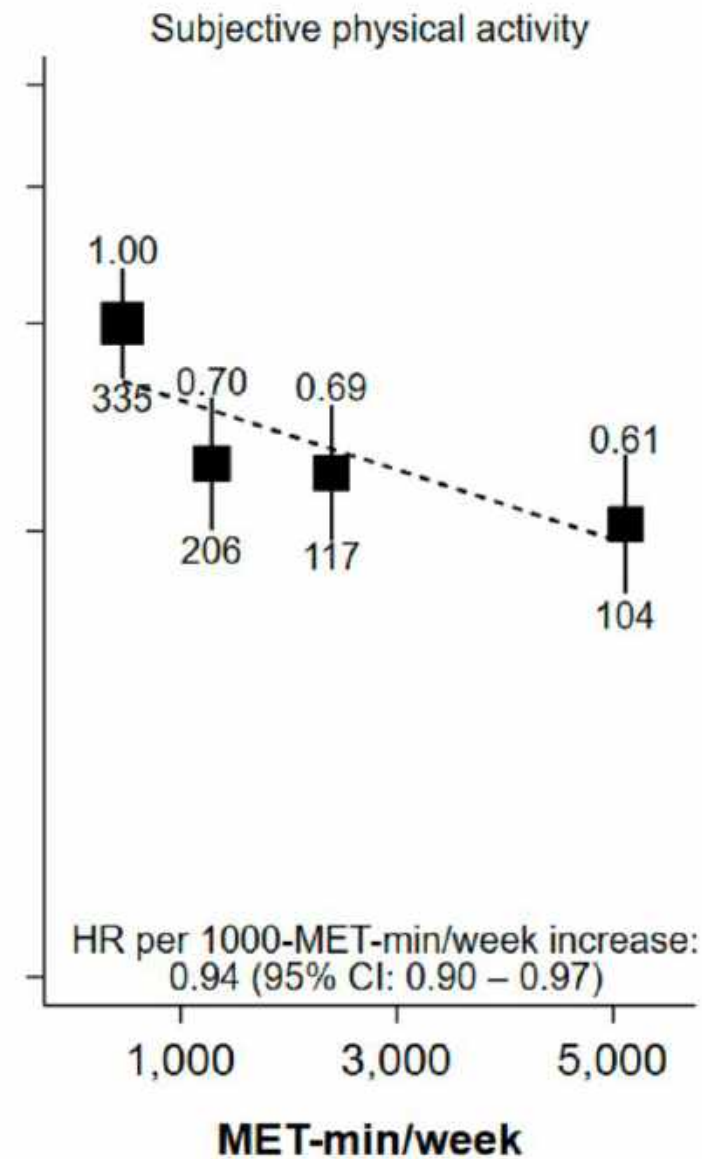
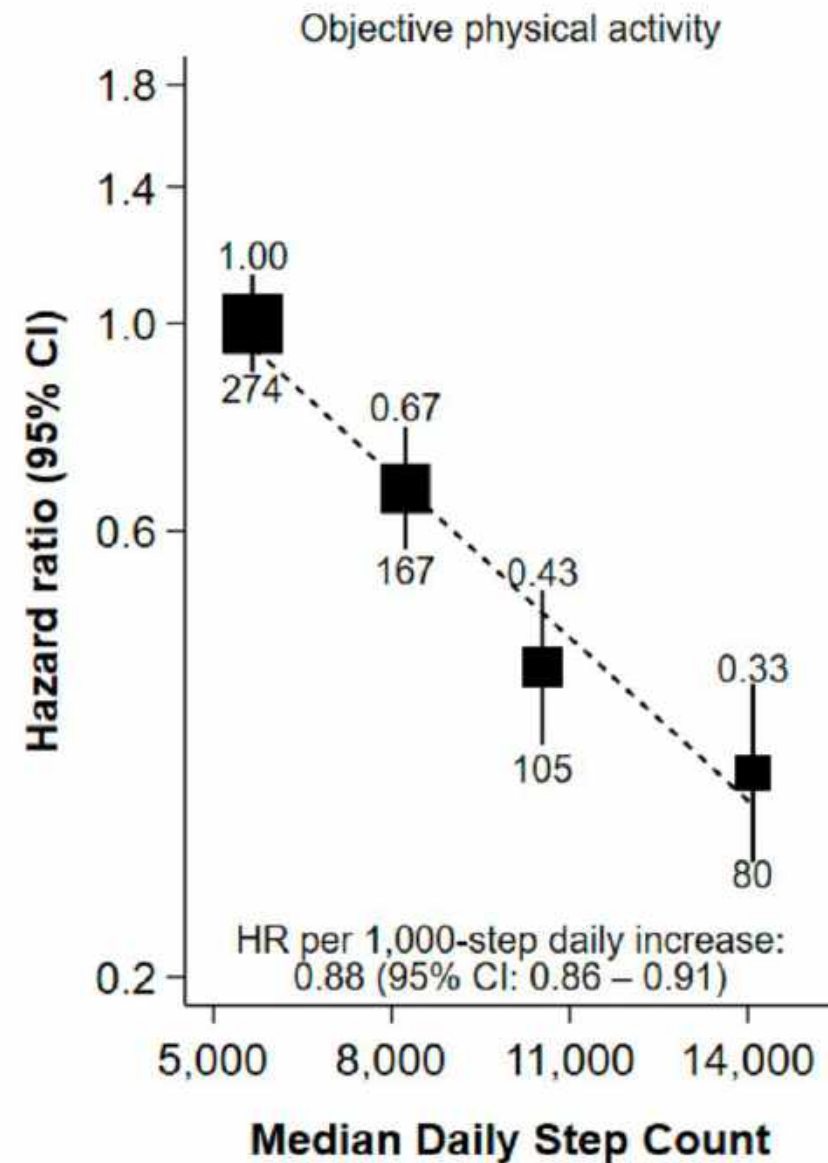
Age and sex adjusted (where appropriate)

Shreves et al. (in preparation)

Associations with cardiometabolic disease - device vs. self-report

7.9 years follow-up

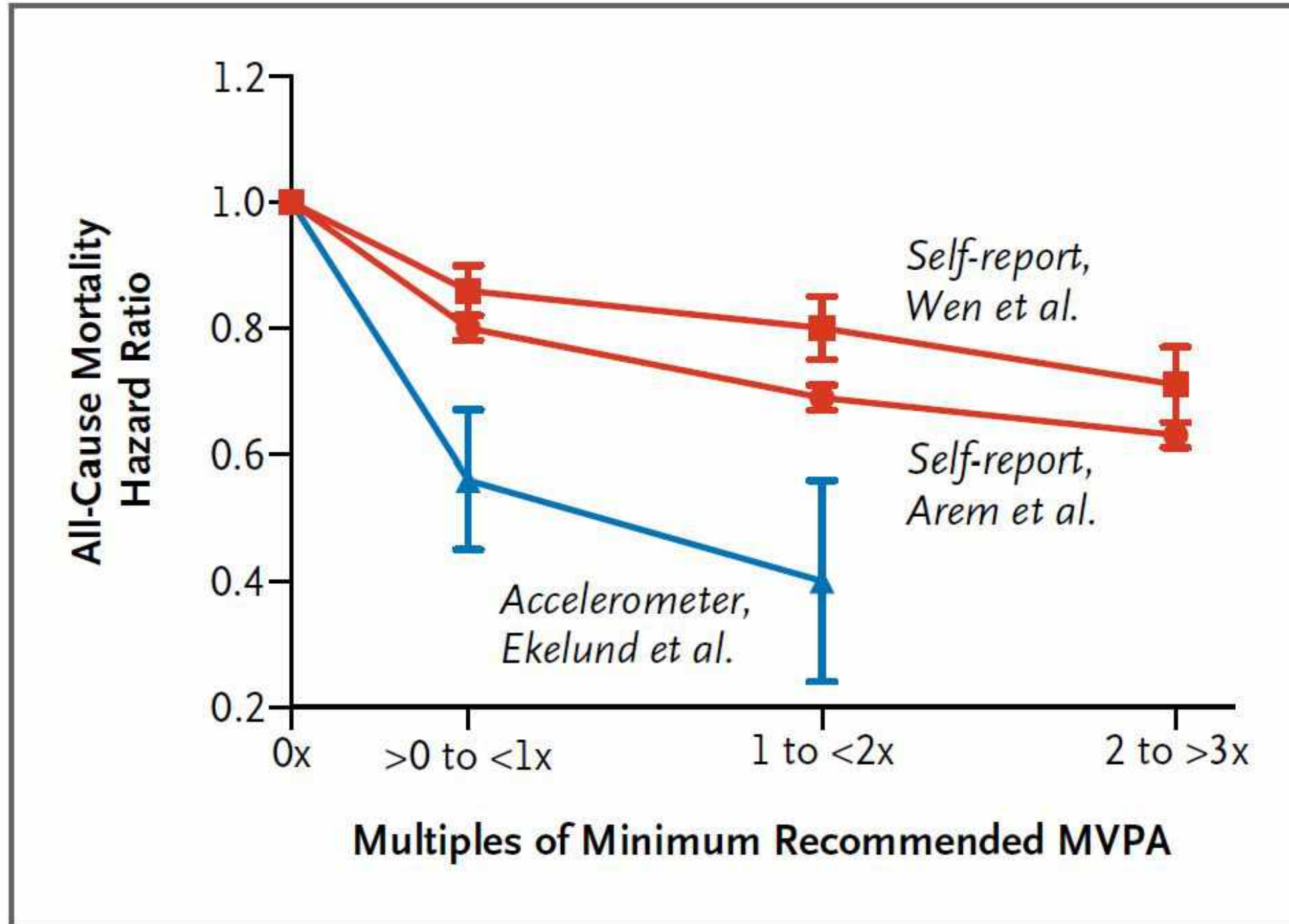
762 Non-alcoholic fatty liver disease events



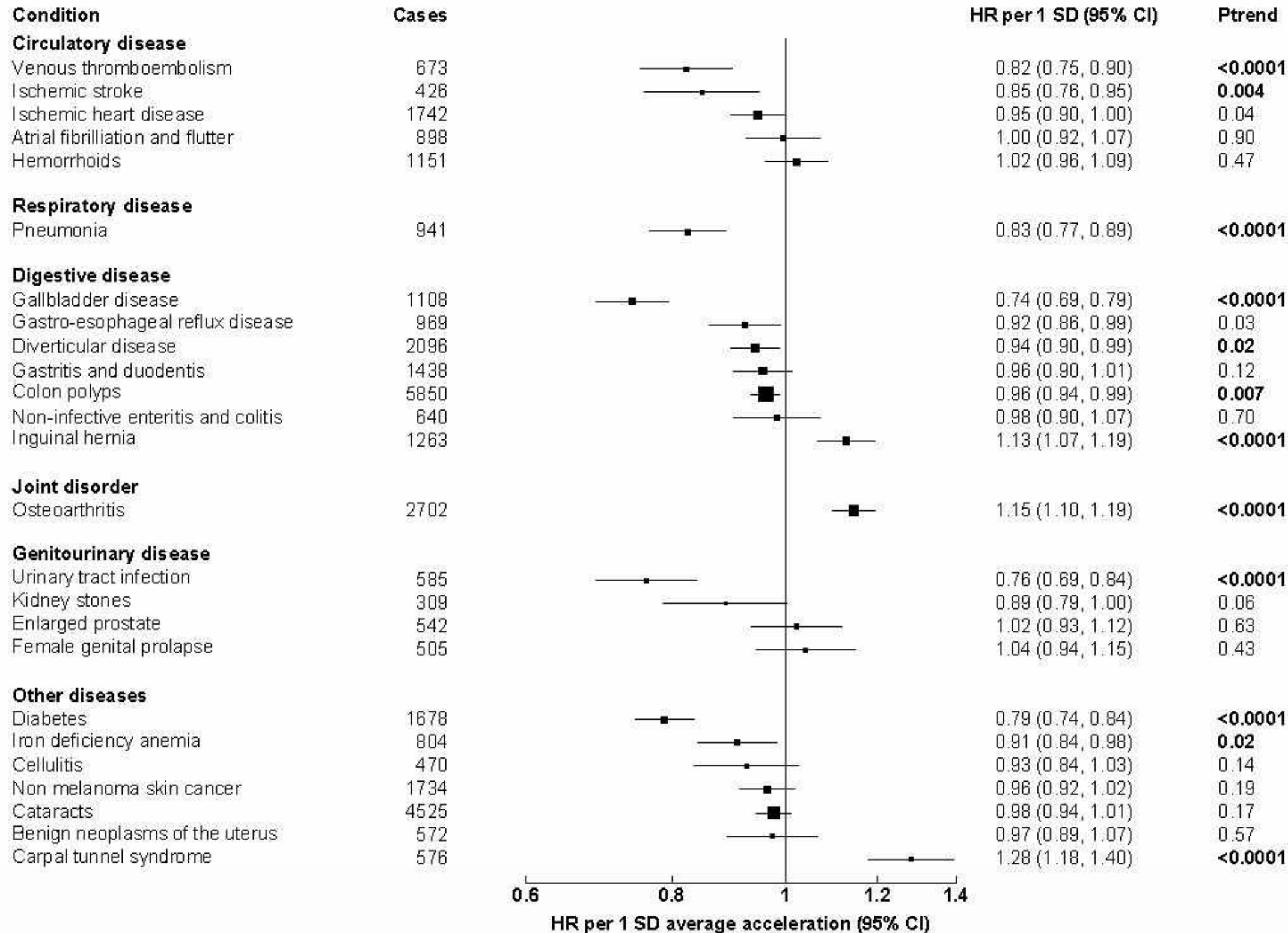
Among
76,464 participants
with both objective
and subjective
activity
measurements

Using age as the time scale and adjusted for sex, ethnicity, Townsend deprivation index, educational attainment, alcohol consumption, smoking status, fruit and vegetable consumption and

Associations with all-cause mortality - device vs. self-report



Associations with common non-cancer outcomes



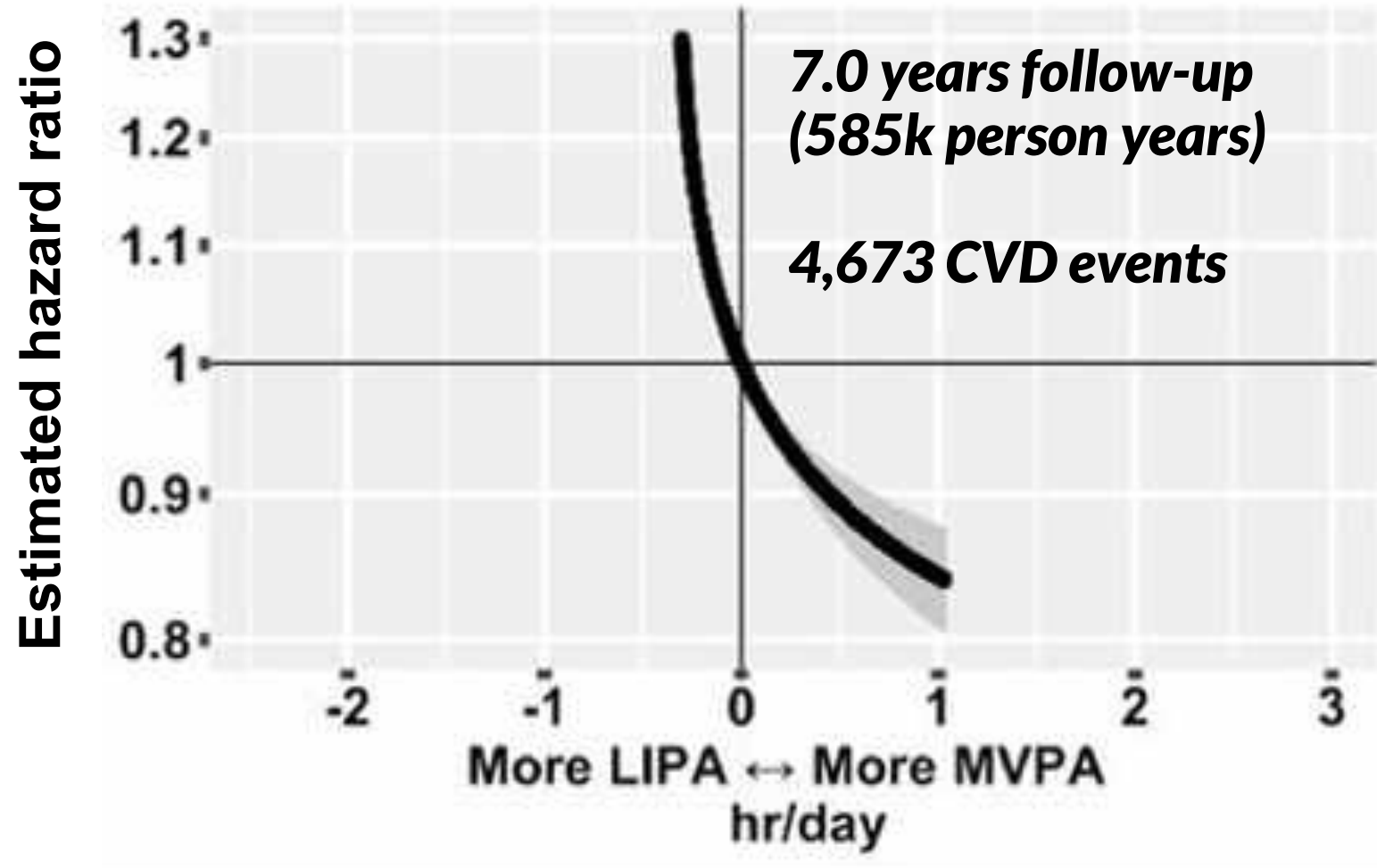
6.8 years follow-up

New insights into activity intensity

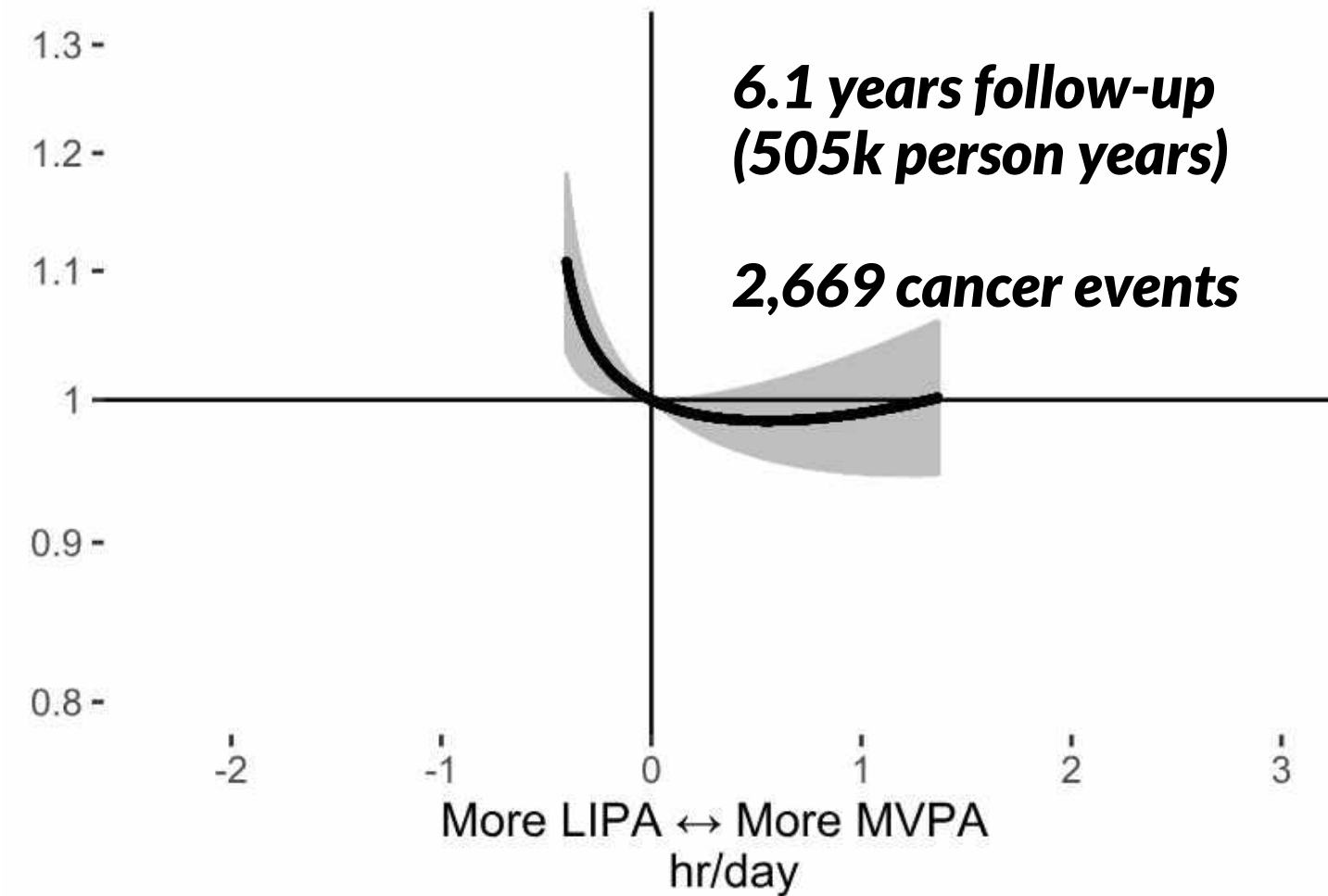
Hazard ratios for incident disease associated with balance between physical behaviours in >86,000 UK Biobank participants



Cardiovascular disease



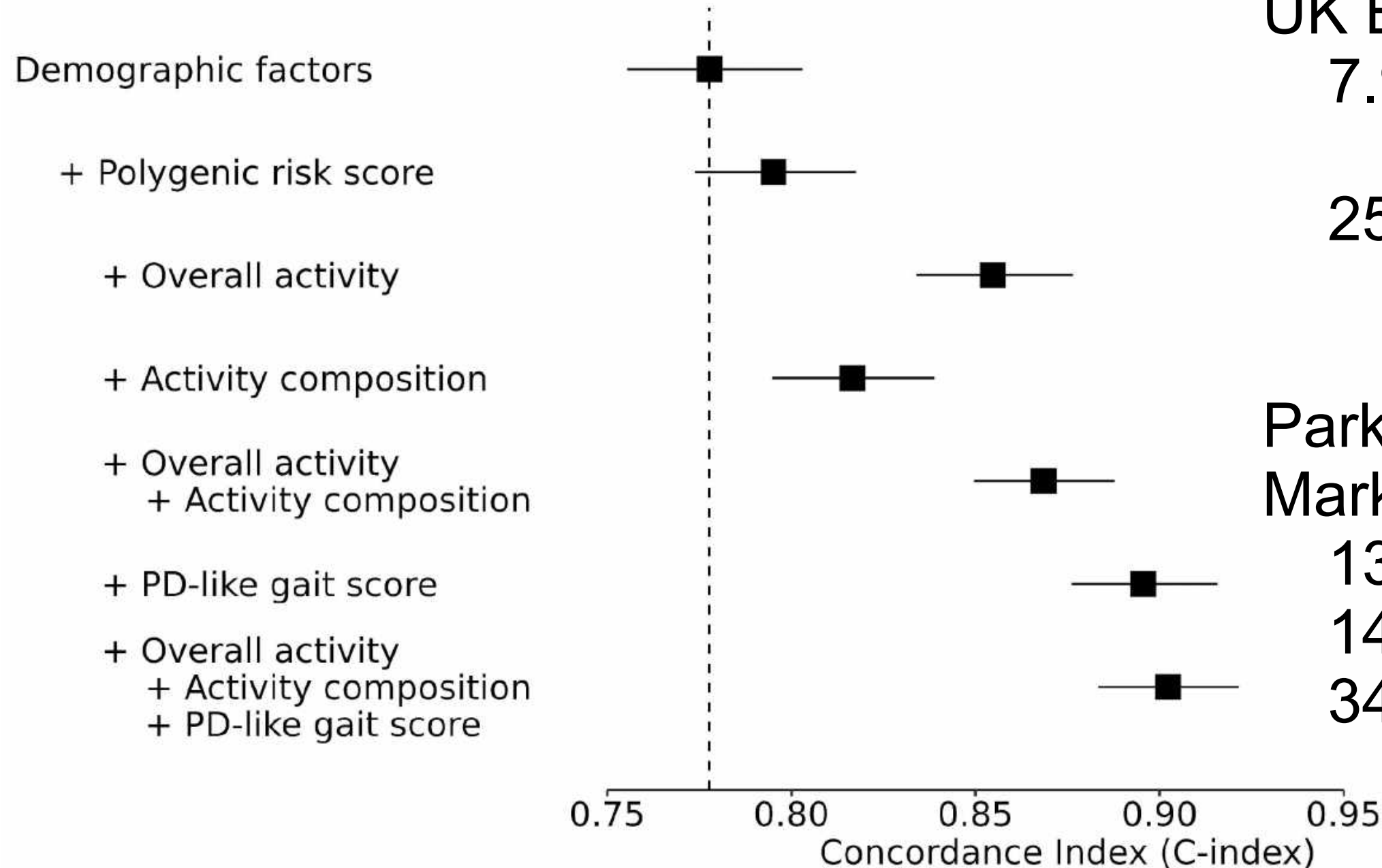
Cancer



Predicting Parkinson's Disease using gait signatures



Concordance Index with 95% Confidence Intervals



UK Biobank:
7.9 years follow-up

259 PD events

Parkinson's Progression
Markers Initiative:

133 Diagnosed PD

147 Prodromal PD

34 Healthy Controls

The UK is falling behind on commercial wearables – All of Us Cohort, USA

**All
of Us**
RESEARCH PROGRAM

Taking at least
8,200 steps a day
may **lower risk for:**



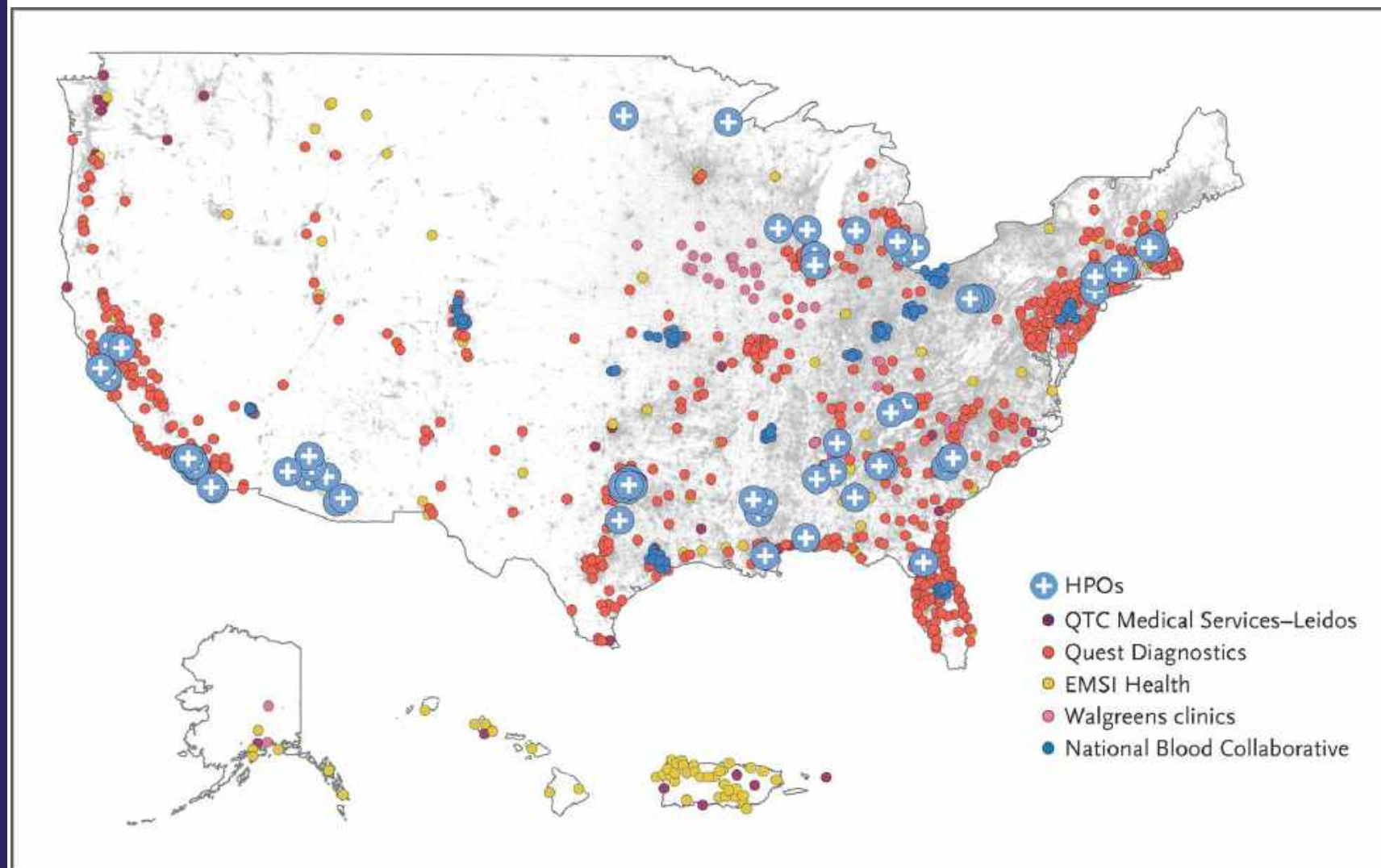
8,200 Steps

Depression*
Sleep Apnea
Acid Reflux

Obesity
Diabetes
High Blood Pressure

*Thanks to the data that *All of Us* participants share, this study found new links between steps and a lower risk for depression, sleep apnea, and acid reflux.

Reference: Master, H., Annis, J., Huang, S. et al. Association of step counts over time with the risk of chronic disease in the *All of Us* Research Program. *Nat Med* (2022).



Digital Footprints Data Integration:

Reflections from experience of linking shopping data

Dr. Anya Skatova
UKRI Fellow
Digital Footprints Lab
IEU/Population Health Sciences
University of Bristol

Scoping review of using shopping data for health research: Problems for Health Research



Burgess et al., in press



Missing context

Actual consumption,
household composition



Missing health information

No knowledge of actual
diagnoses, treatment



Biases

Sampling and
demographic



Data quality

Card not always scanned;
data can be sparse



Shopping across stores

Missing other retailers,
out-of-home consumption



Ethical and

acceptability issues

Data is not collected for
research

Digital Footprints Data in Longitudinal Population Studies



Stage 1: Data linkage

- ALSPAC
- Acceptability
- Ethical and legal basis
- Linkage infrastructure
- Data management

Stage 2: Validating data

- Sampling biases
- Measurement error
- Validating patterns in the data

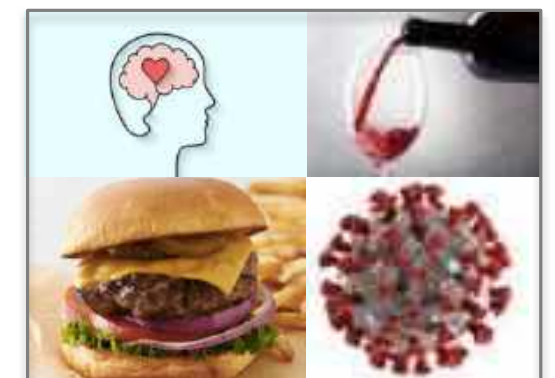
Stage 3: Data Access & Research

- Reproductive health
- Nutrition & lifestyle
- Respiratory illness
- Self-medication

Longitudinal Digital Footprints Data



Population Health



Other Digital Footprints Data



Linking DF data into LPS



Stage 1: Participant acceptability

- Acceptability
- Ethical and legal basis
- Expectations

Stage 2: Data linkage

- Communicate with third parties
- Linkage infrastructure
- Data management

Stage 3: Data Quality

- Sampling biases
- Measurement error
- Validating patterns in the data

Stage 4: Data Access & Research

- Secure
- Ethically compliant
- TRE's



244 participants (63% female)

658,375 items purchased

Between 2013 and 2024

82 categories, 1002 subcategories

55,176 unique products

Our research into the potential of **personal data donation** has shown that >50% of people in the UK are willing to donate their transactional data to health research

– but only if it is in a **trusted**, *transparent* and *secure* manner.



THE CONVERSATION

Academic input, journalistic chair

Donate your data – how your digital footprint can be used for the public good

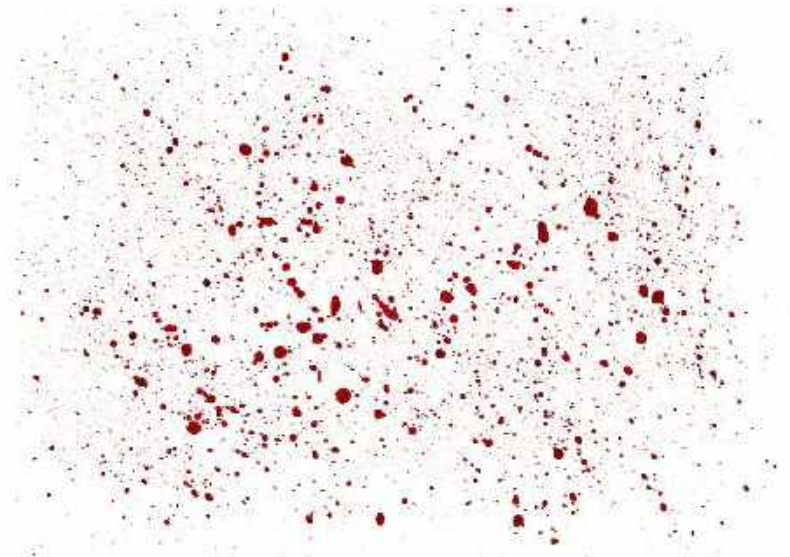
Published: January 12, 2015 2:51 pm GMT

Anya Skatova

Research Fellow, Horizon Digital Economy Research Institute, University of Nottingham

James Goulding

Research Fellow, University of Nottingham



Donate data like blood, and we can look for answers in the patterns we find.
nomad7ra/Shutterstock

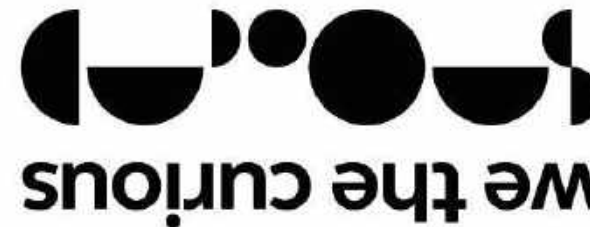


An acceptability and governance foundation for linking participant retailer loyalty card records to UK longitudinal population studies

- **A report** on 5 partnering LPS participants' acceptability of linking shopping data.
- **A PPIE toolkit** including co-created video or/and animation explainers, informative infographics and terminology informed by PPIE insights, enabling effective engagement with participants; protocols to study acceptability and feasibility of using shopping data in LPS through surveys and focus groups;
- A **generalisable ethical and legal framework** for shopping history data linkage within LPS acting as a blueprint to govern Smart Data linkage with LPS/utilising UK LLC. A **training webinar and open access materials** of how to use this framework in practice.

Shopping trolley secrets: from Children of the 90s to We The Curious

- More than public engagement
- We listen to what public says about our research
- We learn which research projects are more important for the public
- We have an amazing opportunity to discuss ethical issues around using shopping data for research in an open and non-university based environment
- This helps us to shape our research practices



By artist Sneha Uplekar
www.microdragons.co.uk



Session 3: Discussion Prompts

- What are the major challenges in integrating internet search data with other health and social data types?
- What technical solutions are needed to facilitate the secure and efficient linkage of diverse datasets?
- How can we ensure the privacy and security of consumer-held data during the integration process?
- What practical steps can be taken to encourage researchers to utilize consumer-held data, including Google data?

Session 3: Summary

Considerations:

Moving Beyond Silos: Integrating internet search data with health records, wearables, and more.

Complex Linkage: Tackling challenges of matching and connecting disparate data types securely.

Technical Solutions: Exploring advanced methods like secure multi-party computation and federated learning.

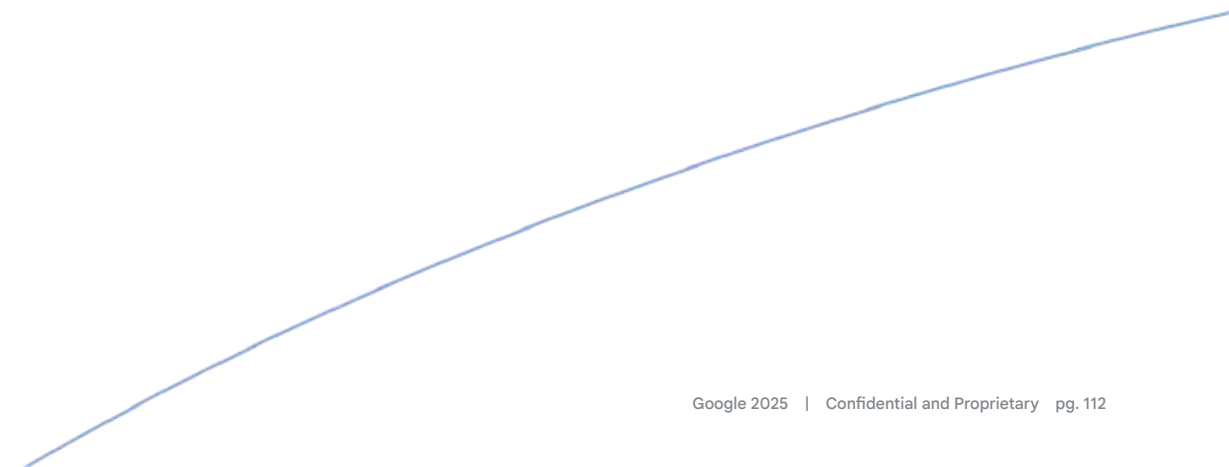
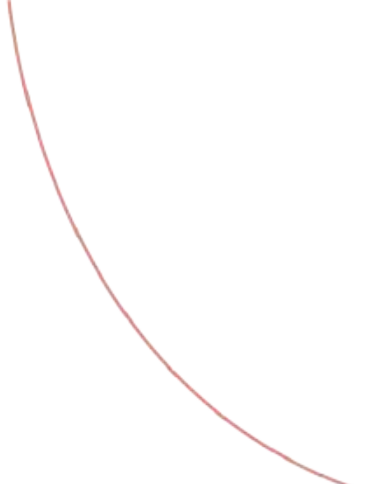
Scalability: Discussing how to enable research at a truly transformative scale

Session 4: Challenges for Research



Suzanne Scott

Public Awareness & Consent



"If I searched for someone else, even it could be anything, I wouldn't like to share you know that sort of information because I wouldn't know that have they gave that consent?"

"Well, if it wasn't health-related... I-I'd question why you need to know that."

"It can be interaction with government, er, websites, can be for my tax files. [...] I wouldn't like such sensitive data to sneak in in my internet use history and shared with, eh, irrelevant people."

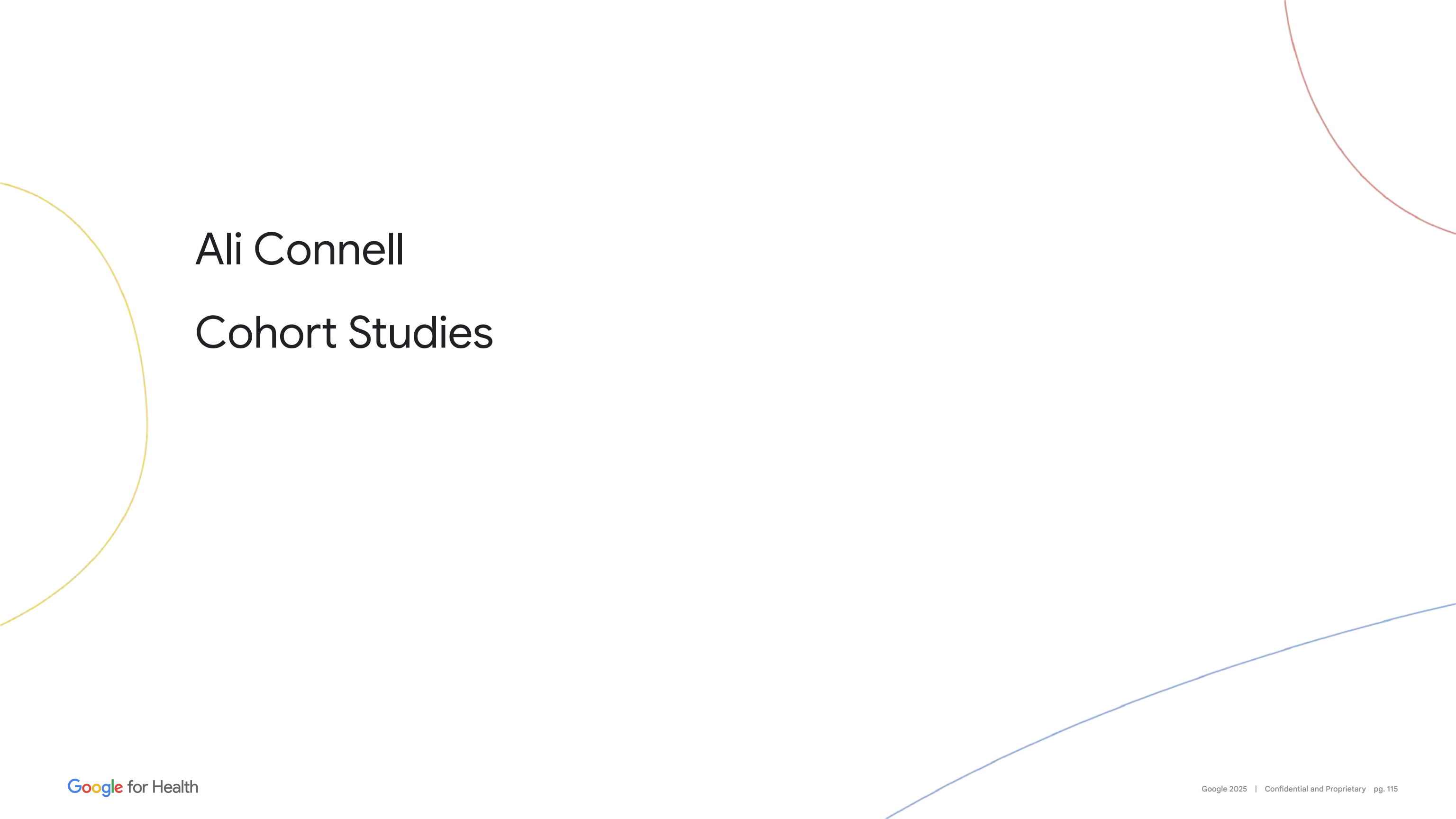
"It's too new to me, I didn't know about Google TakeOut and I didn't know there was such a thing as a Google account."

"I find, personally, find it intrusive. Although I know I'm aware of the low value of this internet use history in my case, eh, I'm not inclined eh, to share it with anyone else."

"I don't mind sharing it as long as it's easy for me to actually share it."

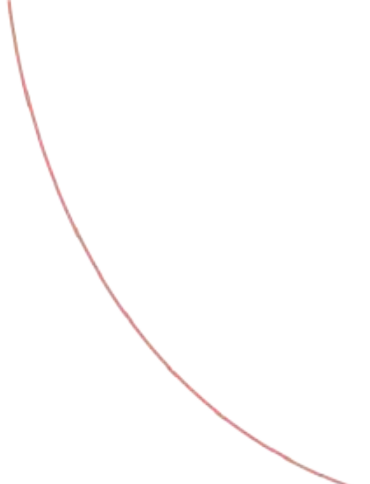
Participant recommendations for future studies

Trust & Transparency	• Provide clear and accessible information about the study and data use. • Use NHS or university branding to build trust. • Ensure participants have control, including opt-out options. • Maintain communication with participants with a clear point of contact for queries. • Include clear information on the research team members including name and credentials.
Privacy, Data Control & Digital Boundaries	• Implement a robust filtering system to isolate health-related data and exclude non-health-related content (e.g., finances, politics, school information). • Reassurance regarding anonymity, data storage, data protection and confidentiality.
Burden of Effort and Digital Literacy	• Offer simple, step-by-step guidance on how to share search history. • Include visual aids or videos to support less digitally confident users. • Provide technical assistance if needed.
Innovation & Societal Impact	• Emphasise the societal benefits, such as improving early diagnosis. • Highlight how the research could help others. • Reassure participants that their contribution is meaningful and valued.



Ali Connell

Cohort Studies



Tim Chico

Study Design

Session 4: Discussion Prompts

- How can we improve public awareness and understanding of research using internet search data to facilitate informed consent?
- What strategies can be used to effectively recruit and consent participants for studies using internet search data?
- How can we best integrate internet search data into existing population cohort studies?
- What are the most effective study designs (prospective vs. retrospective) for research using internet search data?
- What infrastructure, governance is needed?

Session 4: Summary

Considerations:

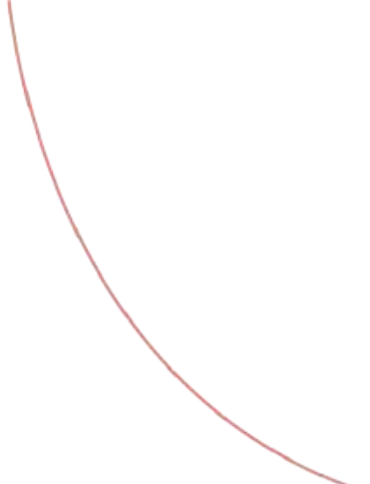
Participant Engagement: Clear, transparent consent and recruitment.

Study Design: Adapting methodologies for novel data sources (prospective vs. retrospective).

Public Trust: Building understanding and confidence in how data is used.

Cohort Integration: Opportunities to enrich existing long-term studies.

Building trust is as important as the data



Session 5: Looking Ahead

Insight to Initiative

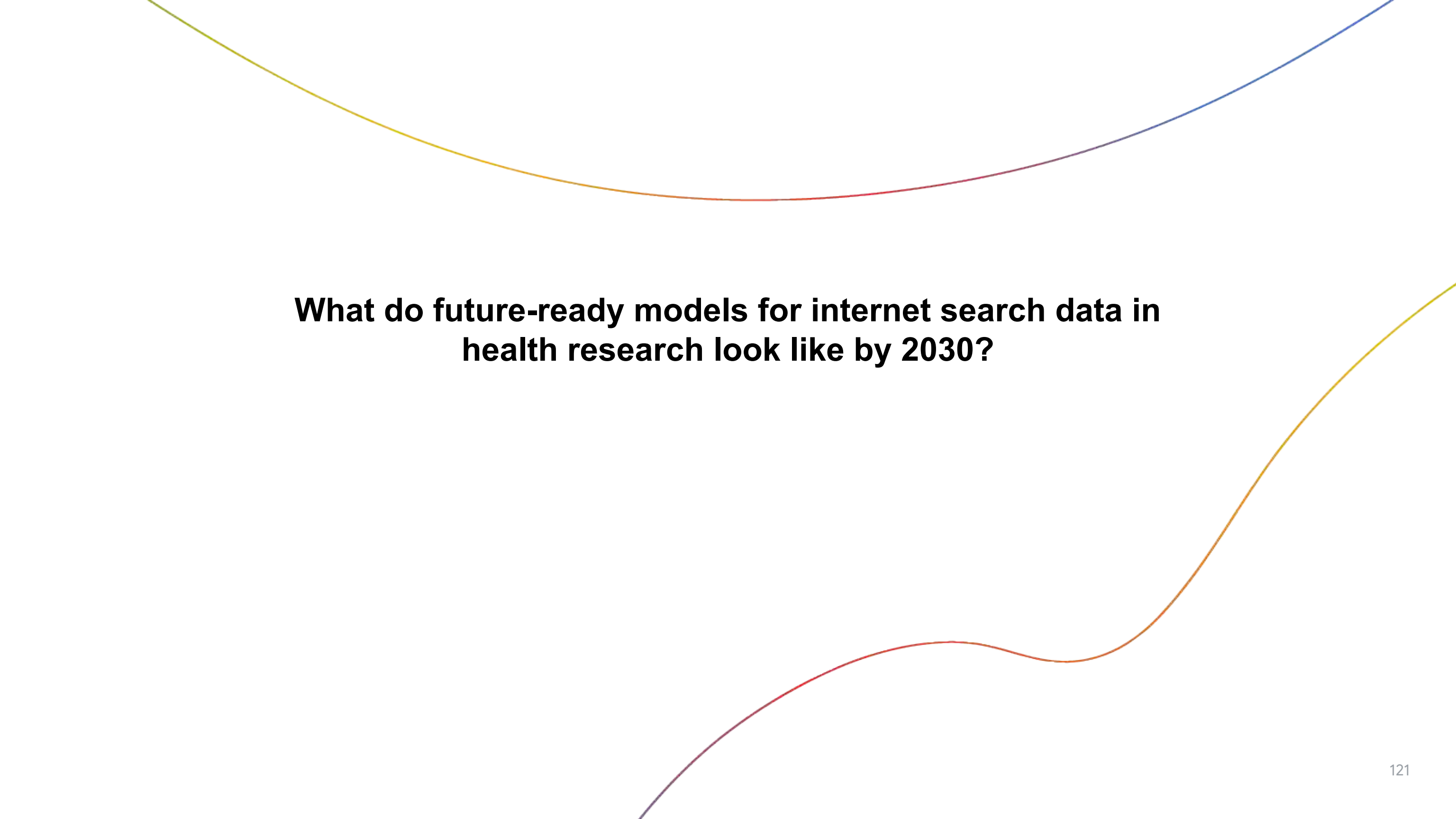


Turning Signal into Strategy

Welcome to our synthesis session

Objective: Capture, prioritize, and act on the most promising ideas from the day

Dr. Matthew Thompson & Eboney White



What do future-ready models for internet search data in health research look like by 2030?

Cross-Session Themes

1. Search data = feasible for behavioral signal.
2. Legal and ethical frameworks are lagging behind innovation.
3. Integration is possible but fragile.
4. Public trust depends on transparency and longitudinal consent.
5. Researcher tooling and data literacy are essential.

Cross-Session Barriers

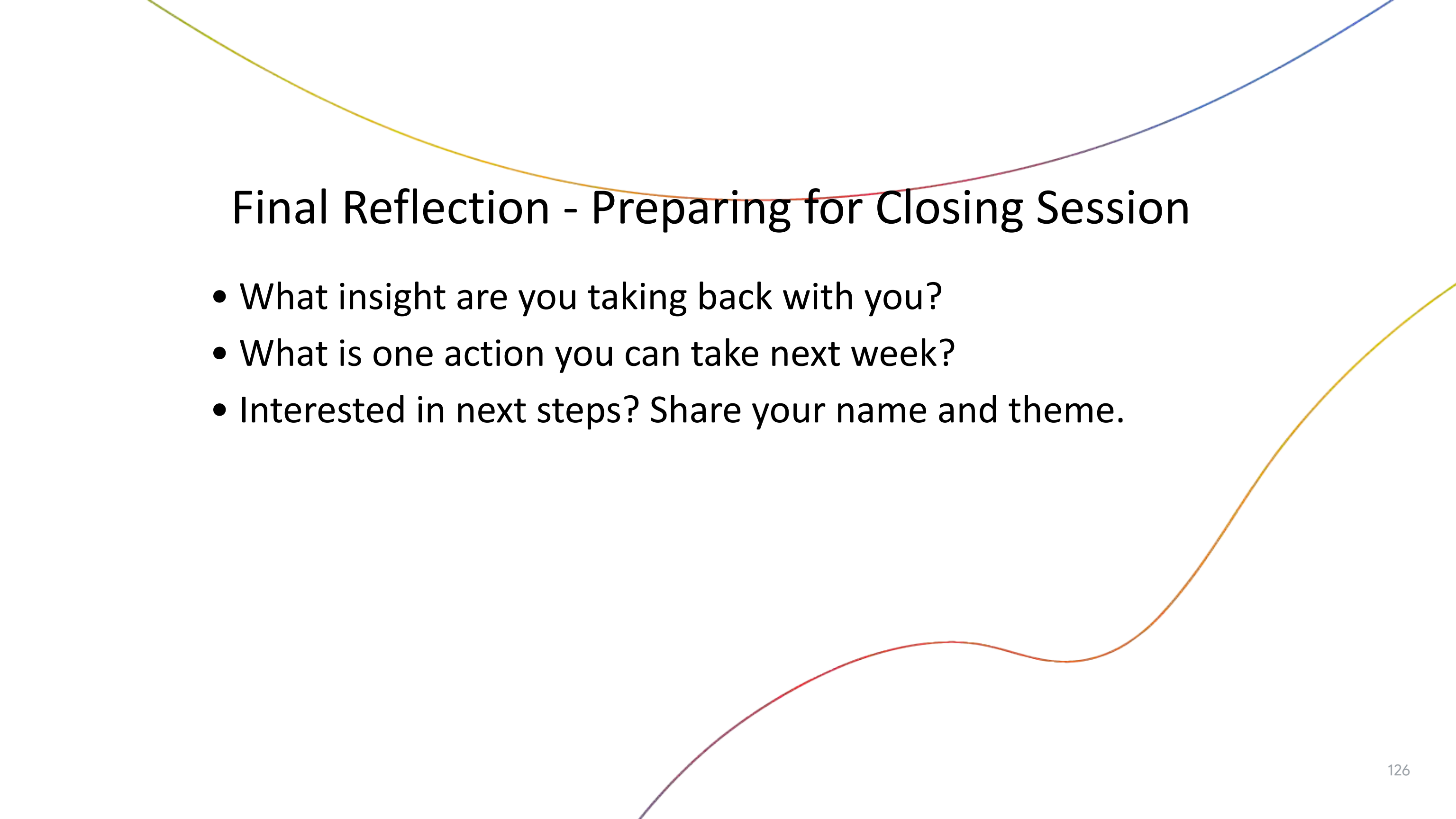
- Legitimacy gap in traditional research settings
- Low public awareness of data donation
- Ethics processes not designed for behavioral data
- Technical fragmentation across systems
- Limited funding mechanisms for emerging data use cases

High Level Group Prompts

- What initiatives should emerge from this convening?
- Who should lead it?
- What data, partnerships, or tools are needed?
- What would success look like in 6–12 months?

Sharing & Prioritization

- Each group shares: One idea, one barrier, one action.
- Attendees vote: What resonates most?
- Look for convergence and momentum.



Final Reflection - Preparing for Closing Session

- What insight are you taking back with you?
- What is one action you can take next week?
- Interested in next steps? Share your name and theme.

Work Session Groups

Group 1: Bridging Data Science, Law, and Public Health	Group 2: Exploring AI, Epidemiology, and Data Governance	Group 3: Integrating Social Data, Cancer Informatics, and Research Strategy	Group 4: Addressing Data Portability, Health Data Epidemiology, and Clinical Research	Group 5: Examining Internet Geography, Research Challenges, and Funding
Jessica Bell	Dan Lewer	Luke Sloan	Tom Fish	Emmanouil Tranos
Aiden Doherty	Agnieszka Scott	James Flanagan	Eva Morris	Suzanne Scott
Jeanelle De Gruchy	Tim Chico	Tarek Al Baghal	Sarah Devaney	Rushil Ranchod
Goran Nenadic	Urszula Pawlicka-Deger	Talisia Quallo	Anya Skatova	Richard Graham
Naomi Herz	David Zendle	Charles Marshall	Tom Fish	Garth Funston

Group 1: Prompts

Bridging Data Science, Law, and Public Health

- What do we not know in health? Could search data help us to fill knowledge voids?
- How can data trusts or similar legal frameworks be adapted to facilitate responsible data sharing and portability for health research using internet search data in the EU and UK?
- From a data science perspective, what are the major challenges in combining internet search data with other health and social data types, and how can legal frameworks support these challenges?
- What innovative funding models or partnerships could incentivize cross-disciplinary research in this area, aligning legal, ethical, and data science needs?
- How do we ensure public awareness campaigns about research using internet search data?

Group 2 Prompts

Exploring AI, Epidemiology, and Data Governance

- What existing studies have successfully utilized individual-controlled internet search data, and what were their key findings, particularly where AI has led to significant epidemiological insights or public health improvements?
- How can Google's data portability and API tools be more effectively used by researchers and further optimized to facilitate AI-driven analysis of internet search data while maintaining strong data governance and privacy?
- What best practices or emerging technologies, including technical solutions for secure and efficient linkage of diverse datasets, can ensure the security and integrity of linked datasets analyzed by AI algorithms?
- How can we develop robust, transparent, and ethically sound frameworks for obtaining informed consent for AI-powered research using personal internet search data, and what strategies can be used to effectively recruit and consent participants?
- What specific policy changes or international standards are needed to govern the ethical use of AI in health research involving internet search data and promote data portability and the use of donated personal data for research?
-

Group 3 Prompts

Integrating Social Data, Cancer Informatics, and Research Strategy

- What are the main barriers preventing researchers from using this data currently, and how can social data science techniques enhance cancer informatics research, especially when combined with internet search data?
- What are the key and unique ethical and legal concerns associated with using individual internet search data for research, particularly when linking social data (including social media and internet searches) with sensitive cancer data?
- How can we ensure the privacy and security of consumer-held data during the integration process, and what data governance models are most effective for protecting privacy while allowing for rich, linked data analysis in cancer research?
- How can we best integrate internet search data into existing population cohort studies, particularly those focused on cancer outcomes and behavior?
- How can individuals with lived experience, especially cancer patients and their support networks, be meaningfully involved in shaping the direction of this research and ensuring research findings are relevant and accessible?

Group 4 Prompts

Addressing Data Portability, Health Data Epidemiology, and Clinical Research

- How can internet search data uniquely contribute to improving public health outcomes compared to other data sources, and what are the real-world challenges and successes of data portability initiatives in health research, particularly regarding individual-controlled internet search data?
- What governance frameworks need to be in place to ensure responsible use of donated personal data, and how can we combine data portability solutions with health data epidemiology techniques to analyze large-scale datasets effectively and responsibly?
- What legal and regulatory solutions are most effective for ensuring ethical data sharing and consent in the context of data portability and clinical research, and what practical steps can be taken to encourage researchers to utilize consumer-held data, including Google data?
- How do we balance the need for detailed clinical data with the availability of less structured but potentially insightful internet search data, and what funding opportunities exist or need to be created to support research using individual-controlled internet search data?
- What are the most effective study designs (prospective vs. retrospective) for research using internet search data, and how can data donation initiatives be designed to ethically and effectively contribute to health data epidemiology and clinical research studies?

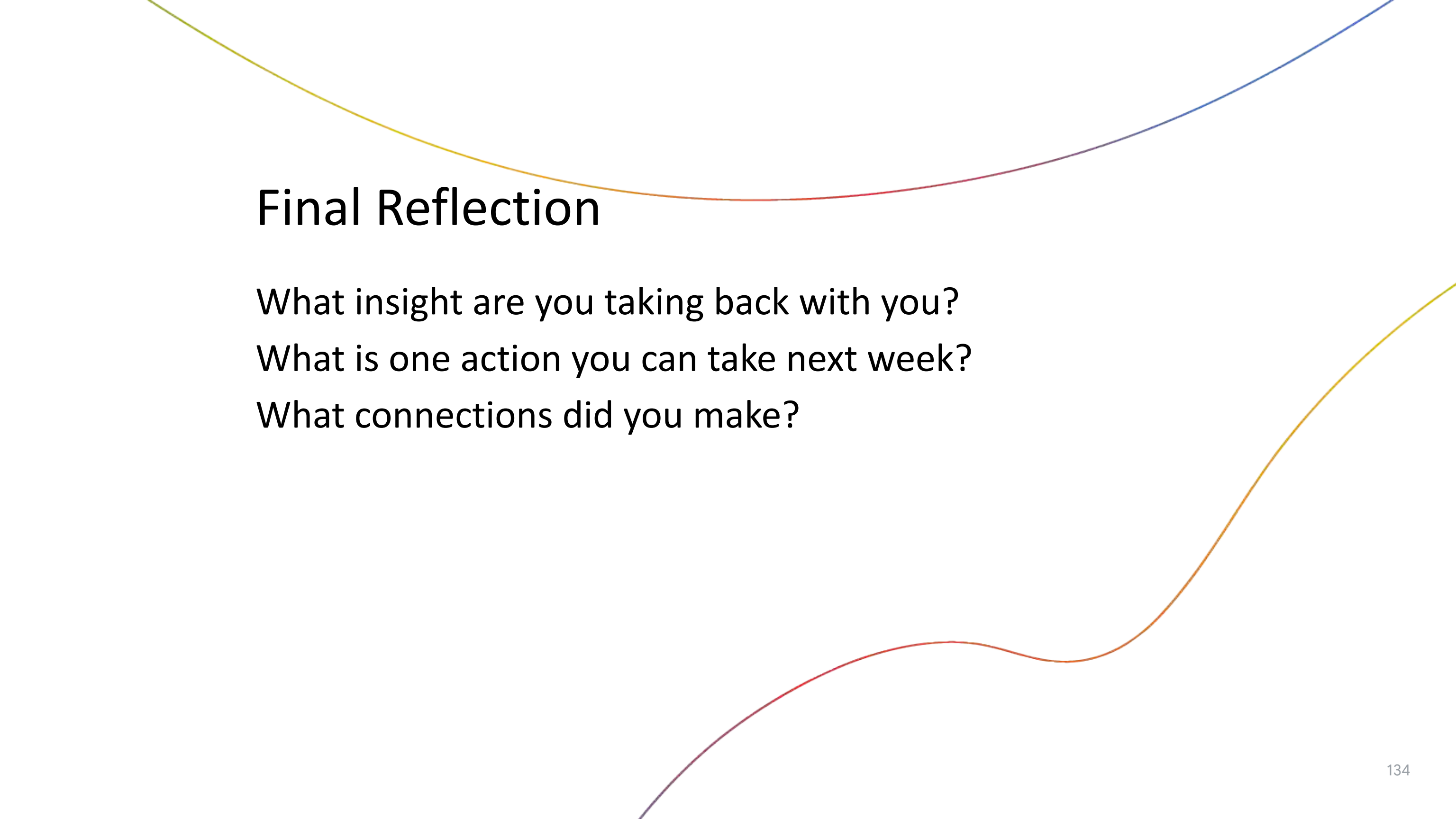
Group 5 Prompts

Examining Internet Geography, Research Challenges, and Funding

- Why are researchers not using these data more extensively, and how can insights from internet geography and web behavior be used to refine health research questions and study designs?
- How can we build public trust in research using internet search data, particularly given concerns about privacy and potential misuse?
- What specific legal barriers or uncertainties are hindering research using internet search data, and how can they be addressed?
- What technical solutions are needed to improve data linkage, portability, and AI-driven analysis of internet search data, and what innovative funding mechanisms can be developed to support interdisciplinary research that combines internet geography, public health, law, and data science?
- What infrastructure and governance models are necessary to support responsible and effective research using internet search data on a large scale?

Session 6: Closing + Wrap Up

Recap & Summary Deck

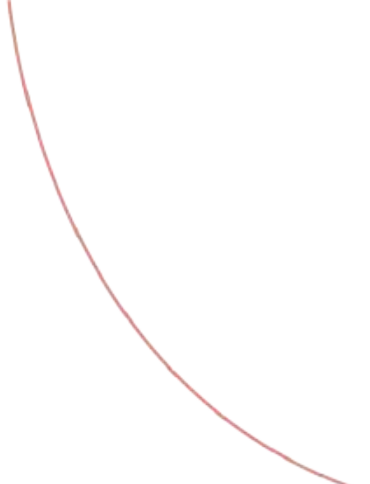


Final Reflection

What insight are you taking back with you?

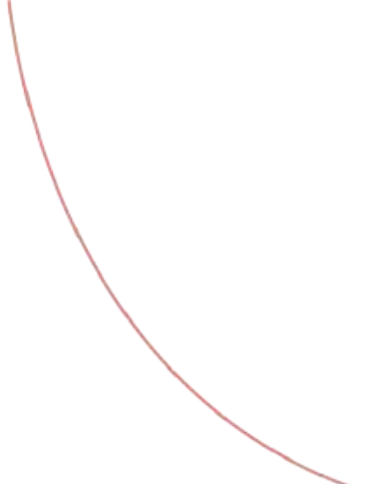
What is one action you can take next week?

What connections did you make?



Emmanouil Tranos

Reflections



Agnieszka Scott

Reflections