SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

RS7: Data Integration

Alexandru Cernat

Thomas O'Toole, Peter Barlow, Natalie Shlomo, Nikos Tzavidis, Joseph Sakshaug

University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick The University of Warwick

Manchester 1824 The University of Manchester

City University of London Est 1894

National Centre for Social Research

LSE The London School of Economics and Political Science

Ulster University

Unil UNIL | Université de Lausanne

Ipsos

KANTAR PUBLIC

What is Data Integration?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Data integration refers to the process of **bringing together information from multiple data sources in a coherent and consistent manner.**

- Data integration makes it possible to examine relationships between factors which might not be visible from any one data source alone.

Research strand 7 of Survey Futures is concerned with how non-survey data can be used to enhance survey data.

What are our Research Themes?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Practitioner Guide 1

Options for integrating non-survey and population survey data.

Practitioner Guide 2

Using integrated non-survey data for evaluating and correcting non-response bias in surveys.

Practitioner Guide 3

Using integrated non-survey data for monitoring and intervening in survey data collection.

University of Essex
University of Southampton
UKRI Economic and Social Research Council

Accessing our work so far

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Practitioner guides: surveyfutures.net/practice-guides/

- Survey Practice Guide 1: Data Integration
- Survey Practice Guide 5: Using integrated non-survey data for evaluating and correcting for non-response
- *Using integrated non-survey data for monitoring and intervening in survey data collection*

Events:

- youtube.com/watch?v=Awjag8x4GiY
- youtube.com/watch?v=VxSC_obOO88
- youtu.be/8Z_g-nb-XG0

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

What's New in Data Integration?

A Scoping Review and Reporting Framework

University of Essex · University of Southampton · UKRI Economic and Social Research Council

NCRM National Centre for Research Methods · Office for National Statistics · UCL · Warwick · Manchester · City · National Centre for Social Research · LSE · Ulster University · Unil Université de Lausanne · Ipsos · Kantar Public

Research Questions

Which survey and non-survey data sources are commonly integrated?

Which methods are most frequently used to link survey and non-survey data?

For what research purposes are integrated survey and non-survey data used?

What is the quality of reporting for data integration?

Methods

We conducted a PRISMA Scoping review (Tricco, 2018)

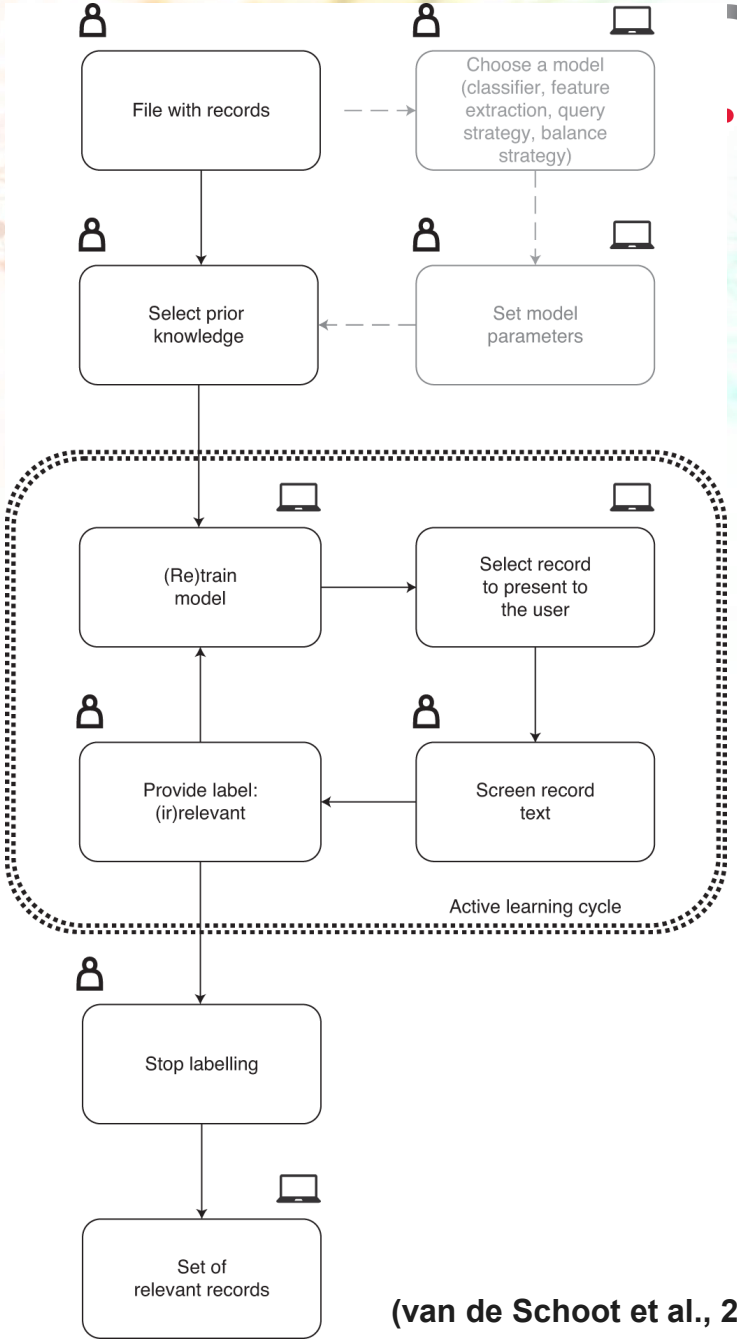
- Ovid (including the Cochrane Library, APAPsycInfo, Embase, Econlit and Medline)
- Web of Science (Core Collection)
- Scopus

Final searches conducted 16/05/2025

- *Screening was calibrated across two coders*

Active learning model (ASReview; van de Schoot et al., 2021)

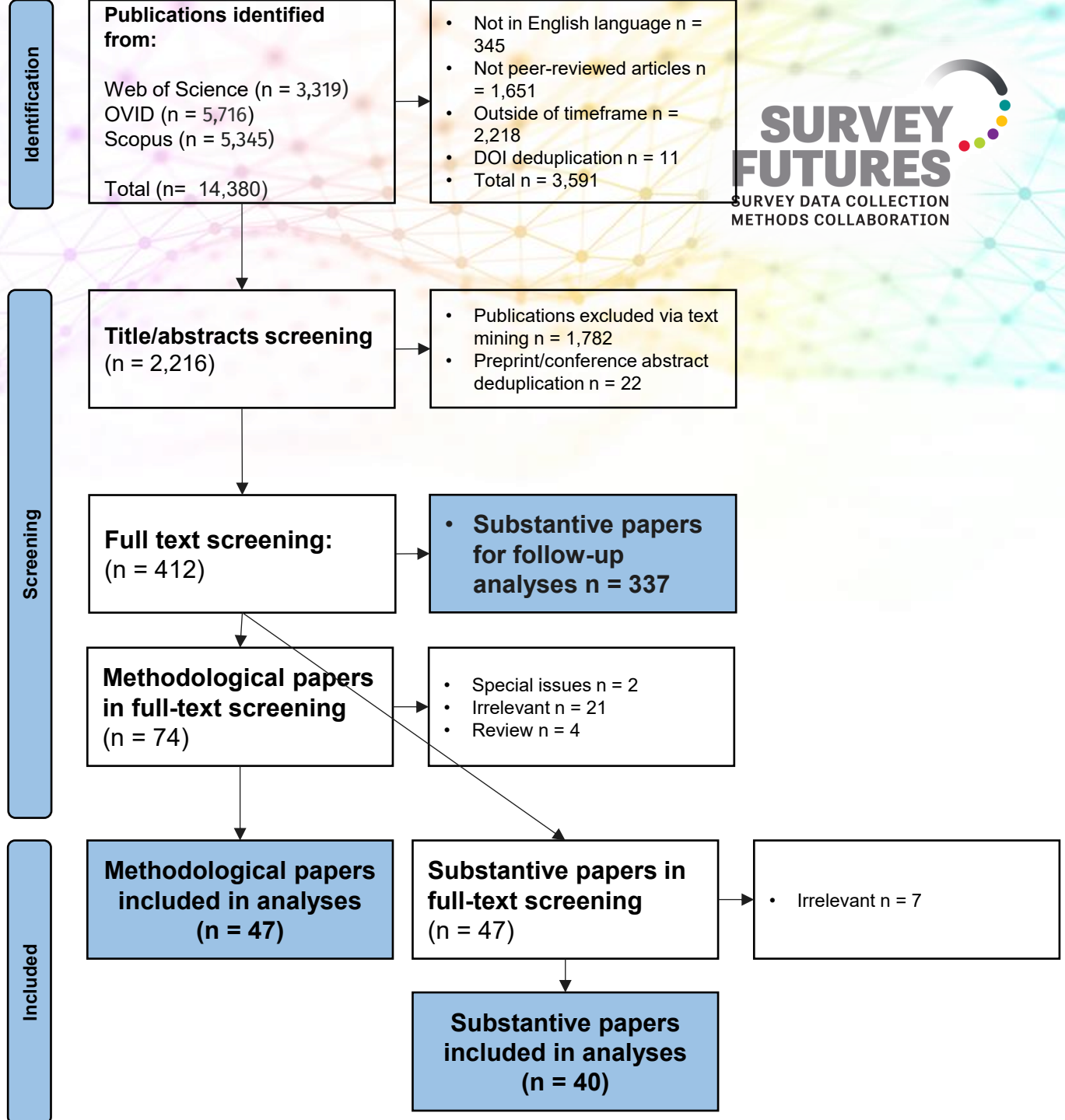
- *Utrecht University's AI Lab "AI-aided Knowledge Discovery".*



(van de Schoot et al., 2021)

University of Essex
University of Southampton
UKRI Economic and Social Research Council

The PRISMA Diagram



SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

University of Essex
University of Southampton
UKRI Economic and Social Research Council

RQ1: Which survey and non-survey data sources are commonly integrated in methods literature?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

	Administrative records	Smart tracker app	Place-based characteristics	Sensor data	Commercial data	Social media data	Total
Panel study	18	4	1		2	1	26
Birth cohort study	3						3
Repeated cross-sectional	13		1	2			16
Cross-sectional	2		1				3
Census	1						1
Total	**37**	**4**	**3**	**2**	**2**	**1**	49

University of Essex
University of Southampton
UKRI Economic and Social Research Council

RQ1: Which survey and non-survey data sources are commonly integrated in substantive literature?

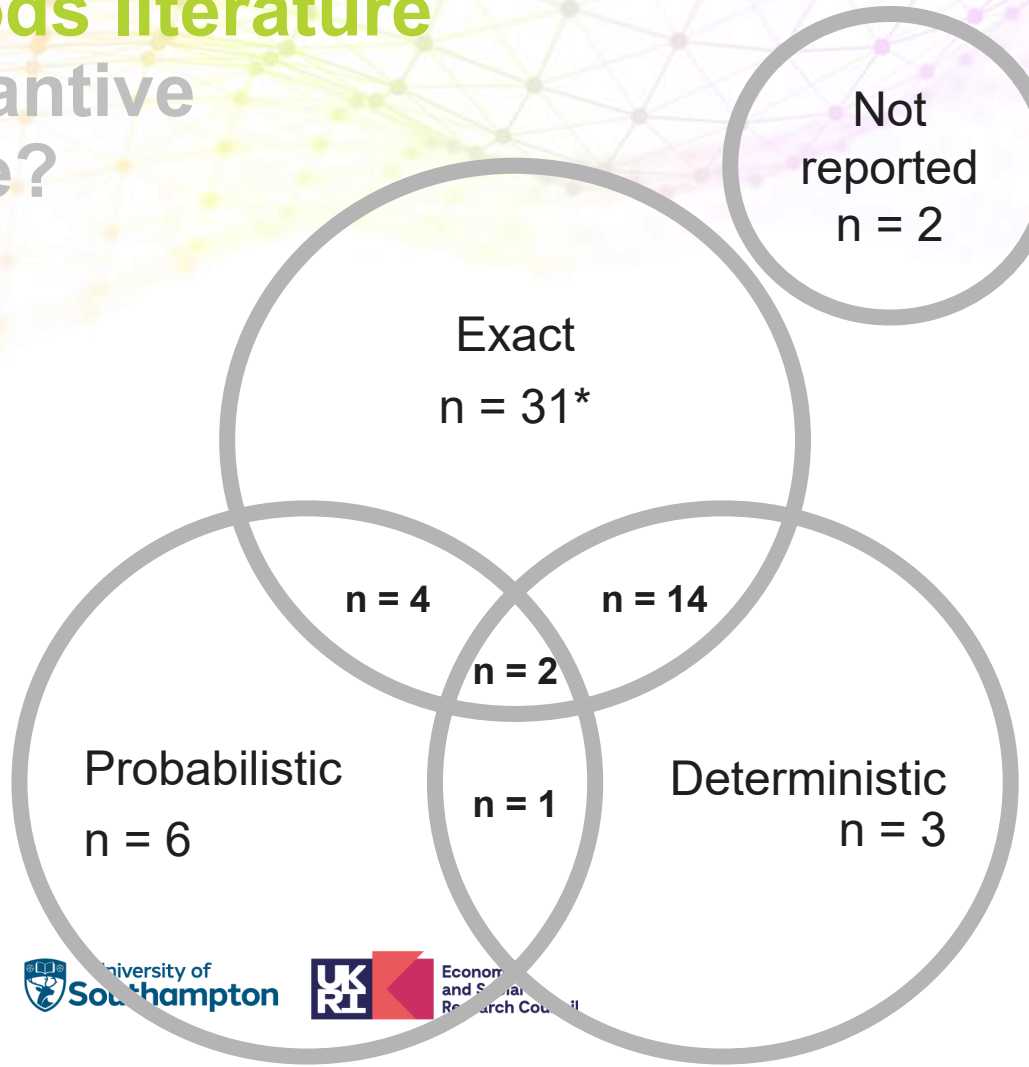
SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

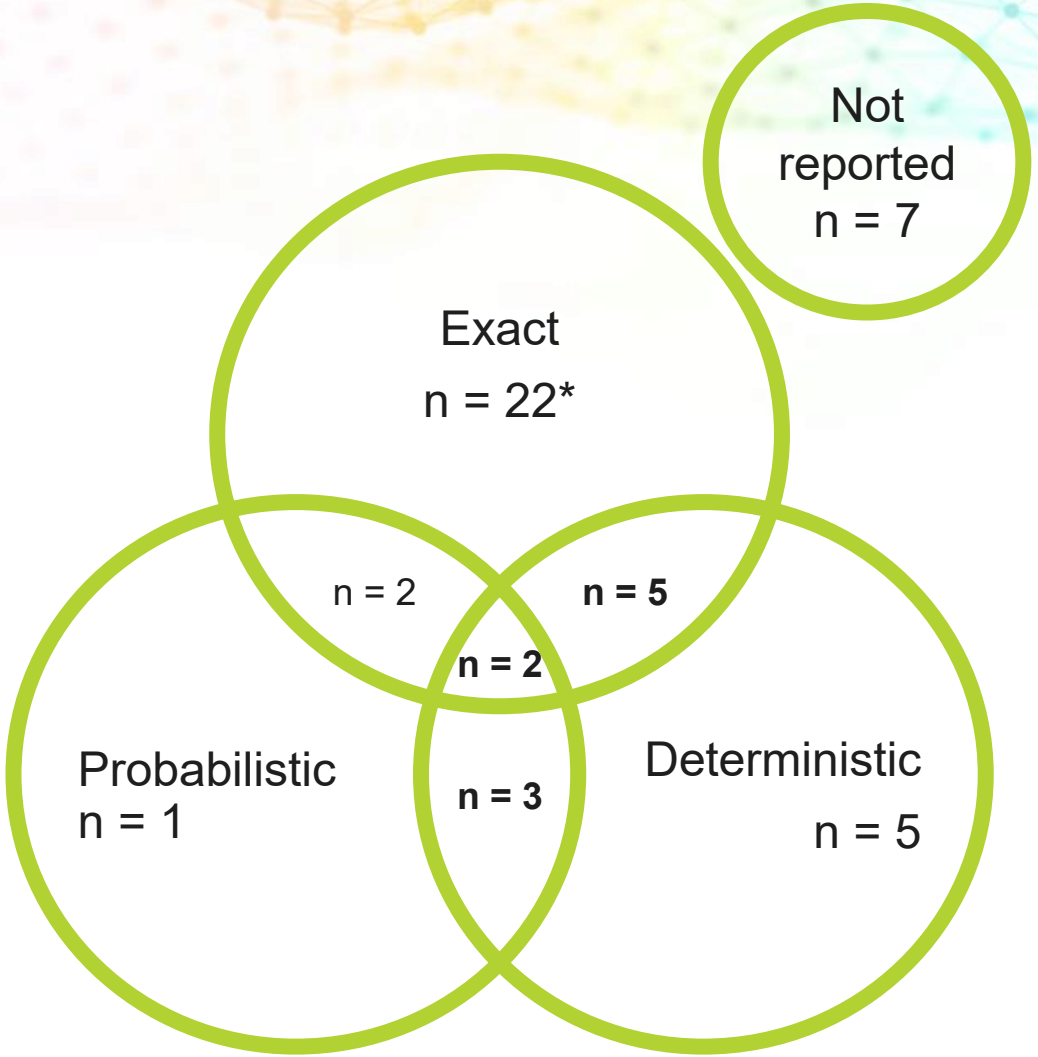
	Administrative records	Smart tracker app	Place-based characteristics	Sensor data	Commercial data	Social media data	Total
Panel study							
Birth cohort study	22						22
Repeated cross-sectional	17	1	4				22
Cross-sectional	7				1		8
Census	28		1				29
Total	**74**	**1**	**5**		**1**		**81**

University of Essex
University of Southampton
UKRI Economic and Social Research Council

RQ2: Which methods are most frequently used to link survey and non-survey data in methods literature in substantive literature?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

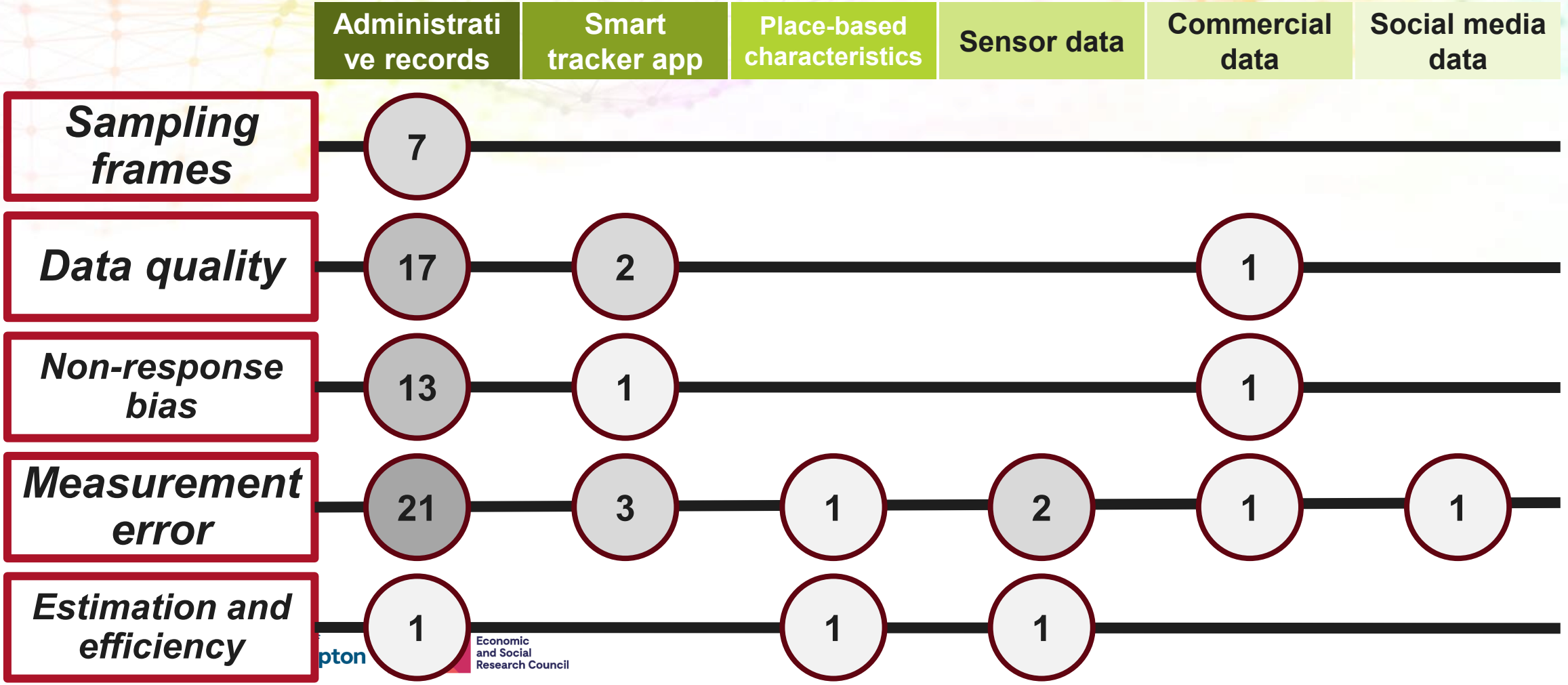




University of Essex
University of Southampton
UKRI Economic and Social Research Council

RQ3: For what research purposes are integrated survey and non-survey data used?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION



pton

Economic and Social Research Council

RQ4: What is the quality of reporting for data integration?

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Linkage and consent details were difficult to find and were inconsistently reported.

40.4% Reported a response rate.

31.9% Reported a consent rate.

55.3% Mentioned consent to linkage.

46.8% Reported a linkage rate.

42.5% Mentioned whether consent was needed.

University of Essex
University of Southampton
UKRI Economic and Social Research Council

Reporting guidelines for Integrated Data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Survey data

(PRICSSA; Seidenberg, Moser & West, 2023)

Non-survey data

- Title
- Type
- Sub-type
- Timeframe
- Population
- Citation
- DOI

Integrated data information

- User guide/technical document
- Citation
- DOI

University of Essex | University of Southampton | UKRI Economic and Social Research Council

Reporting guidelines for Integrated Data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Integration procedure	Integration quality	Integrated data access
• Exact matching • Deterministic matching • Probabilistic matching • Statistical matching • Linkage units	• Linkage consent information • Linkage consent rate • Successful linkage rate • Corrections applied	• Access conditions • Access procedure

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Evaluating Data Quality in the Linked 1970 Birth Cohort Study

University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick

Manchester

City

National Centre for Social Research

LSE

Ulster University

Unil Université de Lausanne

Ipsos

Kantar Public

Research Questions

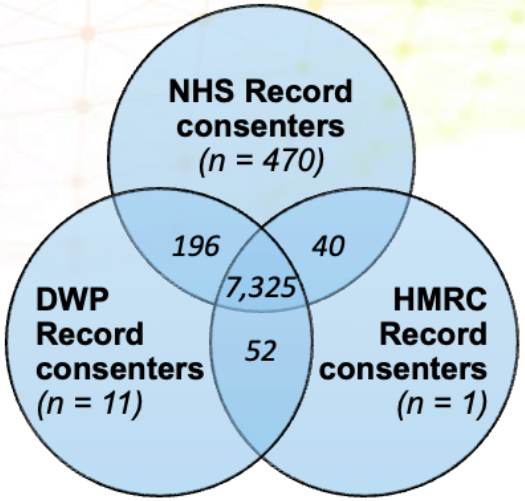
What is the bias introduced by data linkage at each stage?

To what degree does this vary by domain/type of data?

Do differences in sample composition at each stage introduce bias to applied modelling?

Data and selection process

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION



1970 Birth Cohort Study Age 42 (2012)
(n = 9,841)
(RR = 57%)

NHS Record Consenters
(n = 8,033)
(CR = 82%)

DWP Record Consenters
(n = 7,585)
(CR = 77%)

HMRC Record Consenters
(n = 7,419)
(CR = 75%)

HES records successfully linked Age 42 (2012)
(n = 6,099)
(LR = 62%)

SMR records successfully linked Age 42 (2012)
(n = 542)
(LR = 5%)

Non-consent
NHS
(No = 1,299)
(NA = 508)

DWP
(No = 1,746)
(NA = 509)

HMRC
(No = 1,912)
(NA = 509)

Non-linkage
HES
(No = 3,739)
(NA = 2)

University of Essex
University of Southampton
UKRI Economic and Social Research Council

Predictors of consent

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

The strongest predictor of consent was **prior non-response**

Renting was positively associated with consent, as was lower **educational** attainment (GCSE/O-level).

Former smoking status was also positively associated with consent

Several **geographic regions** showed significantly higher odds of consent, particularly the East Midlands, North East, North West, South East, South West, West Midlands, Wales, and Yorkshire & Humber

The same set of variables were found to be associated with almost all linkage types

Selection process by type of linkage

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Dataset	R Indicator	Coefficient of Variation	n
Age 42 Sweep	0.82	0.11	9705
NHS Consenters	0.84	0.10	7957
HES Linkages	0.83	0.10	6041
SMR Linkages	**0.77**	**0.14**	**541**
DWP Consenters	0.84	0.09	7514
HMRC Consenters	0.84	0.09	7352

Note: Response propensities were estimated via binomial logistic regression modelling (Next wave response, as predicted by previous non-response, NHS consent, DWP consent, HMRC consent, sex, Socio-economic classification, household size, house ownership or renting, country of birth, highest educational qualification, marital status, smoking status, economic activity, benefit receipt, general health and government office region.

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Does Additional Contact Information Improve Survey Outcomes?

Hafsteinn Birgir Einarsson, Kim Backström & Alexandru Cernat

University of Essex · University of Southampton · UKRI Economic and Social Research Council

NCRM National Centre for Research Methods · Office for National Statistics · UCL · Warwick · Manchester · City · National Centre for Social Research · LSE · Ulster University · Unil Université de Lausanne · Ipsos · Kantar Public

Research Questions

What is the impact of additional contact information on participation?

How does the composition of the sample change due to extra contact info?

Other ongoing research

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

We investigate the effects of a **change in the Finnish Statistics Act** that granted Statistics Finland access to new contact information from the Finnish Social Insurance Institution.

The additional contact information included

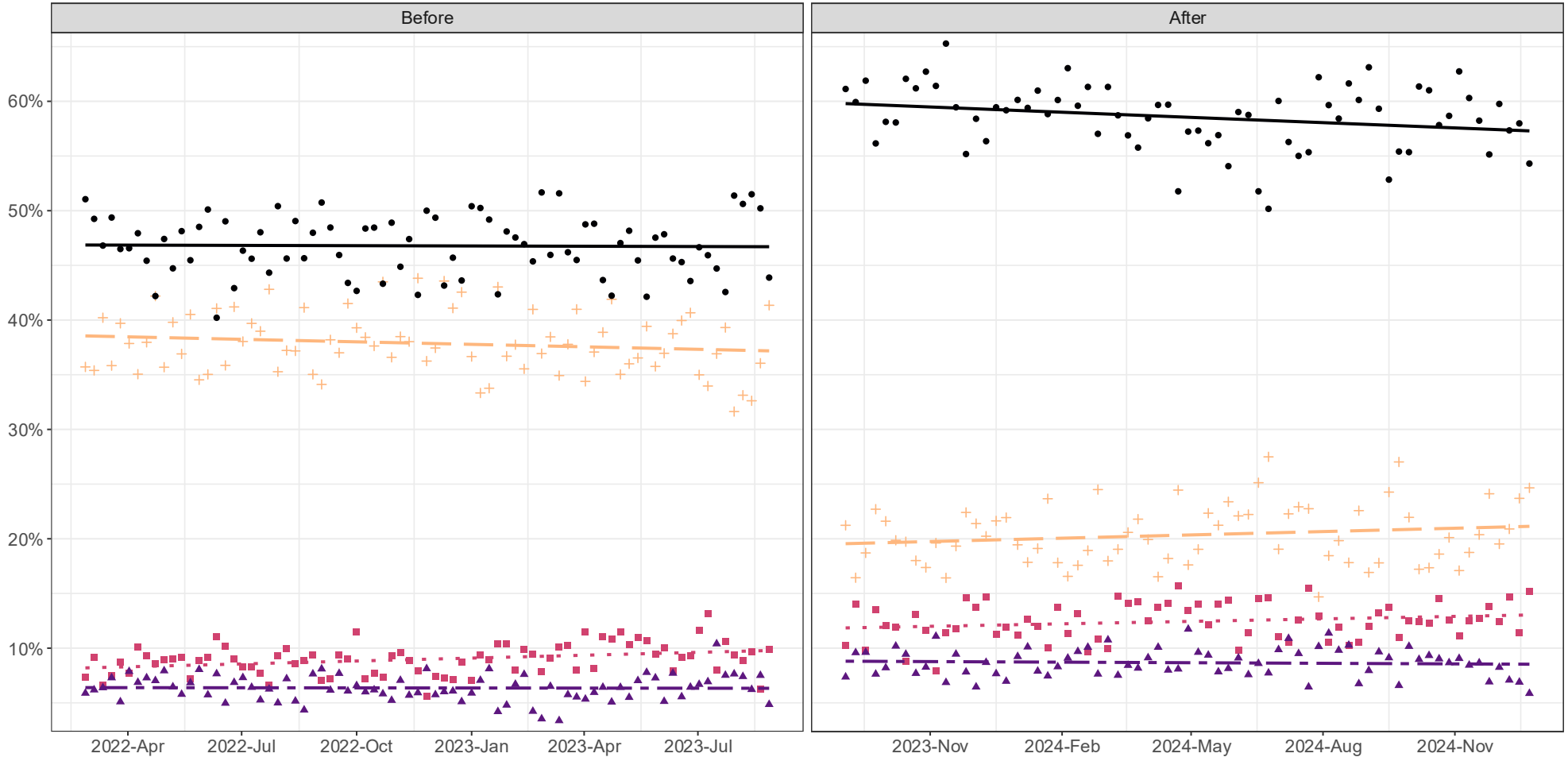
- phone numbers and
- email addresses

We treat **the legislative change as a quasi-experimental design** to investigate how extra contact information affected **response rates and selection bias in the Finnish LFS**

We look at **new samples of ~600 persons randomly selected from the population register each week before and after the policy change** (148 weeks)

Overall participation outcomes

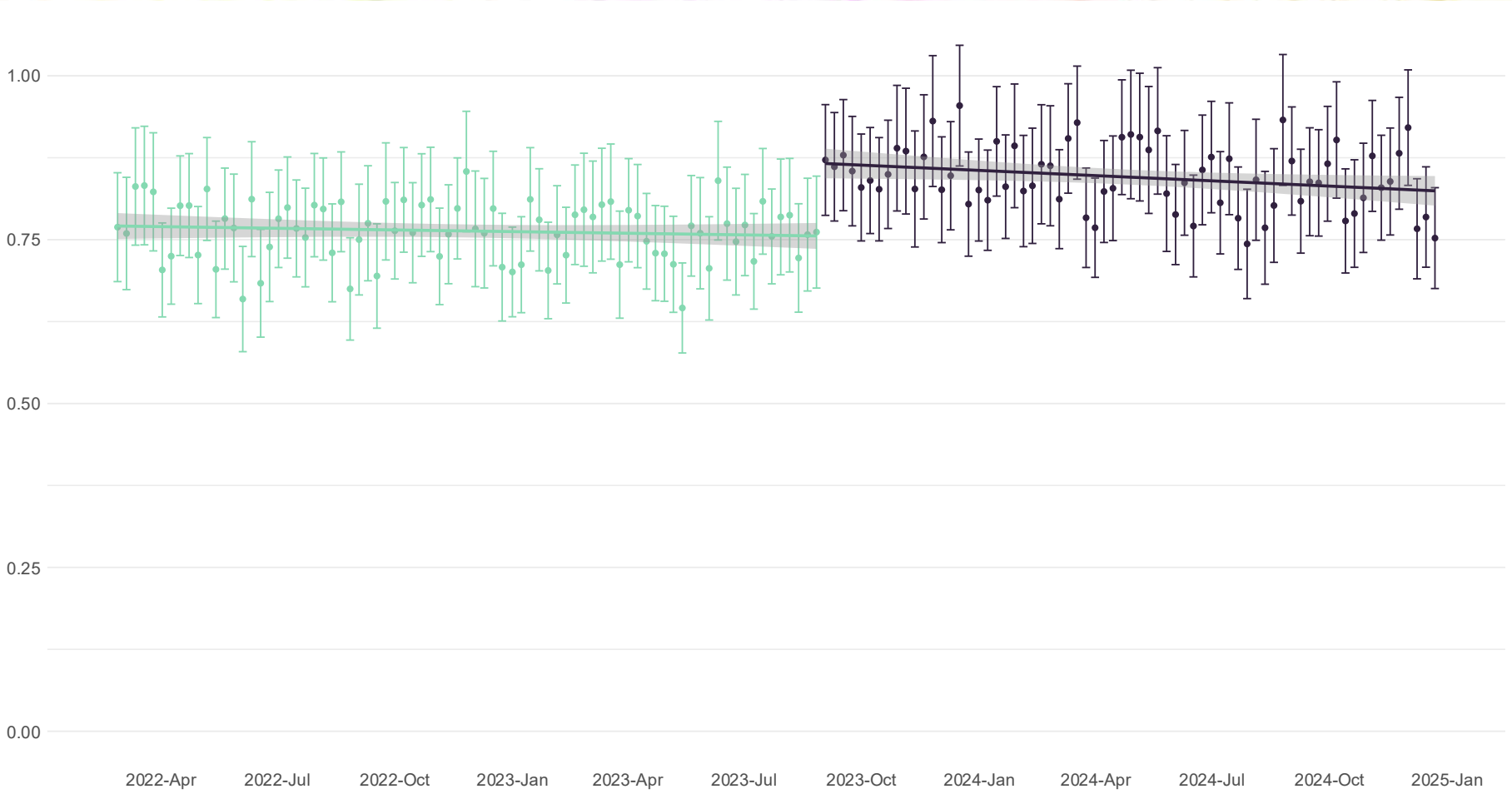
SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION



University of Essex

R-Indicators measuring the representativeness of respondent compositions in FI-LFS

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION



University of Essex
University of Southampton
UKRI Economic and Social Research Council

Partial R indicators show inconsistent changes in selection process

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

University of Essex
University of Southampton
UKRI Economic and Social Research Council

Other ongoing research

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Peter is leading a paper looking at how data at different levels (individuals, HHs, LSOA) predicts non-response in Understanding Society and how important it is to have up to date geographical data

Natalie is working on extending the classical Multi-Frame integration where both x_i and y_i are available in probability and nonprobability samples and same questionnaire applied to both samples

Nikos doing work around integrating survey and geospecial data

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

RS7: Data Integration

Alexandru Cernat

Thomas O'Toole, Peter Barlow, Natalie Shlomo, Nikos Tzavidis, Joseph Sakshaug

University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick The University of Warwick

Manchester 1824 The University of Manchester

City University of London Est 1894

National Centre for Social Research

LSE The London School of Economics and Political Science

Ulster University

Unil UNIL | Université de Lausanne

Ipsos

KANTAR PUBLIC

What search terms were used?

Survey Data	AND	Non-Survey Data	AND	Data Integration
survey* OR "social survey*" OR "household survey*" OR "population representat*" OR population-based OR "nationally representat*" OR "random sampl*" OR "probability sampl*" OR questionnair* OR interview* OR "cross-sectional stud*" OR "longitudinal stud*" OR "cohort stud*"		administrative OR operational OR "routinely collect*" OR "electronic record*" OR geospatial OR geographic OR spatial OR "gridded population*" OR "geographic information system*" OR satellite OR commercial OR "smart app*" OR "digital trace*" OR web-tracking OR "data donation" OR census		"data link*" OR "data integrat*" OR "record link*" OR "probabilistic link*" OR "deterministic link*" OR "data merg*" OR "data fus*" OR datalink* OR "multiple dataset*" OR "multiple data source*" OR "linkage stud*" OR "adaptive survey design*" OR "responsive survey design*" OR "total survey error*" OR "measurement error*"

Economic and Social Research Council

Exclusion criteria

- Not in English language
- Not a peer-reviewed research paper
- Not inside timeframe (2020 – current)
- Not using both survey *and* non-survey data

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

University of Essex
University of Southampton
UKRI Economic and Social Research Council