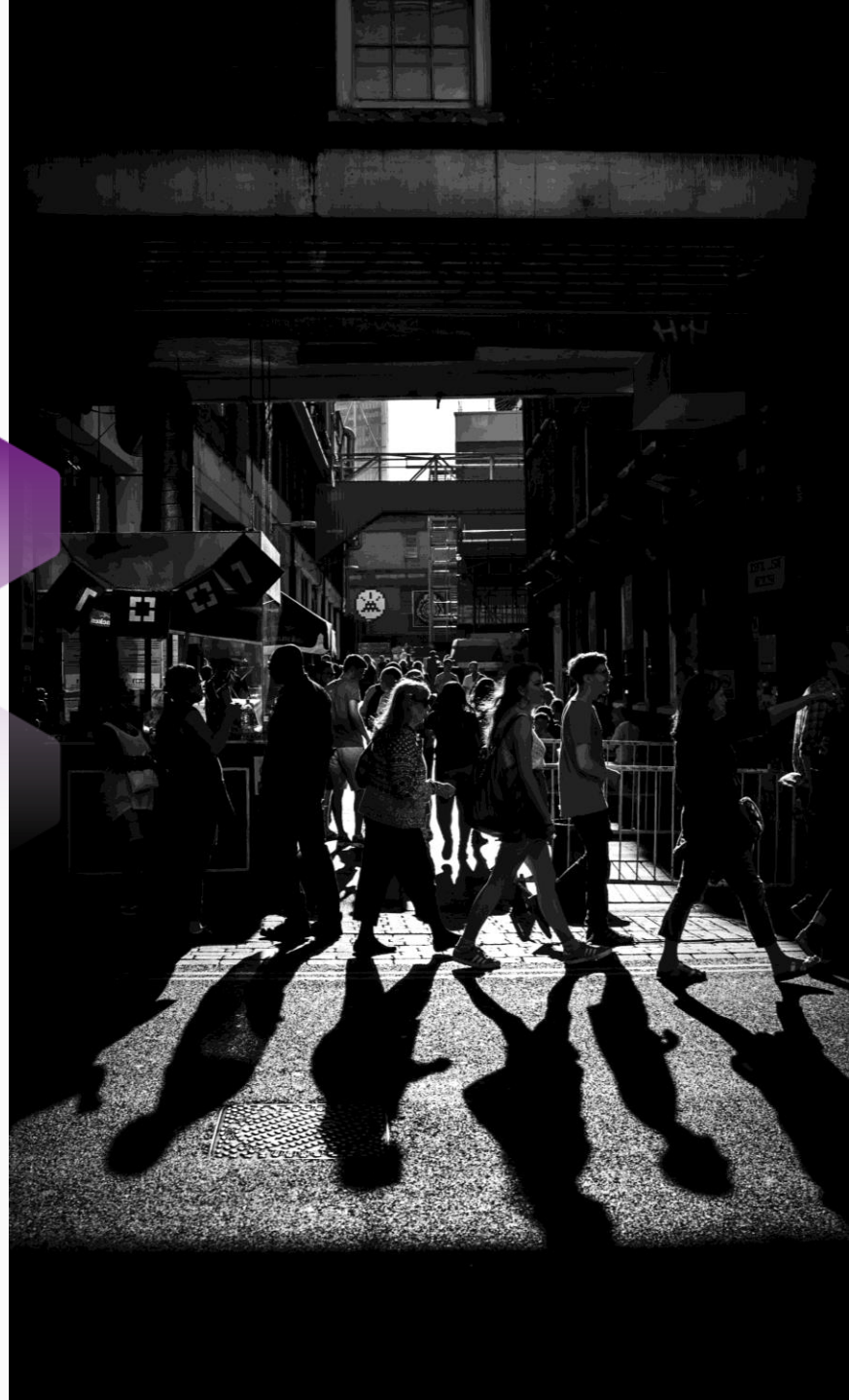
UK Data Service

Patterns in researcher output requests and reasons for refusal

Preliminary findings from the SOCRATES project

Allison Tyler & James Rayner



UKRI Economic and Social Research Council

HDRUK Health Data Research UK

ADRUK Data-driven change

SOCRATES Project

- DARE UK-funded Early Adopter project
- August 2025-October 2026
- Purpose:
  - Test the utility and feasibility of the Automated Checking of Research Outputs (ACRO) tool within a large-scale TRE environment
  - Provide evidence for whether semi-automated SDC checking improves TRE workflow efficiency without compromising data security or compliance
  - Evaluate SATRE specifications against existing accreditation (ISO, DEA)

SOCRATES Project

- DARE UK-funded Early Adopter project
- August 2025-October 2026
- Purpose:
  - **Test the utility and feasibility of the Automated Checking of Research Outputs (ACRO) tool within a large-scale TRE environment**
  - Provide evidence for whether semi-automated SDC checking improves TRE workflow efficiency without compromising data security or compliance
  - Evaluate SATRE specifications against existing accreditation (ISO, DEA)

ACRO Tool

- ACRO is a free and open source tool that supports the semi-automated checking of research outputs for privacy disclosure within secure data environments
  - Developed out of DARE UK Phase 1 Driver Project led by UWE
    - SACRO: Semi-Automated Checking of Research Outputs
  - https://github.com/AI-SDC/ACRO
  - UKRI-funded project in partnership with HDR UK, ADR UK
  - Several DARE UK Early Adopter projects investigating utility of ACRO tool in different TRE environments

ACRO Tool

- Stata, R, python tools that create “safe” tables and figures for researchers
  - Flags potential problems so that researchers can address them before submitting output request
  - Can implement suppression automatically during creation
  - Coding replaces “standard” table/figure development coding
- What ACRO tool can currently do:
  - Cross-tab (frequency tables, descriptive stats, summary stats)
  - Linear Regression (OLS)
  - Logistic Regression
  - Probit Regression
  - Survivability tables
  - Histograms (python only)
  - Survivability plots (python only)

Pandas crosstab

[6]:
```
table = pd.crosstab(df.year, [df.survivor, df.grant_type])
table
```

[6]:

survivor	Dead in 2015		Alive in 2015			
grant_type	G	R	G	N	R	R/G
year						
2010	3	47	12	59	24	8
2011	3	47	12	59	24	8
2012	3	47	12	59	24	8
2013	3	47	12	59	24	8
2014	3	47	12	59	24	8
2015	3	47	12	59	24	8

ACRO crosstab

[7]:
```
safe_table = acro.crosstab(df.year, [df.survivor, df.grant_type])
safe_table
```

```
INFO:acro:get_summary(): fail; threshold: 12 cells may need suppressing;
INFO:acro:outcome_df:
-----------------------------------------------------------------|
survivor   |Dead_in_2015     |Alive_in_2015                      |
grant_type |G            R   |G            N   R   R/G           |
year       |                 |                                   |
-----------------------------------------------------------------|
2010       | threshold;   ok | ok           ok  ok  threshold;   |
2011       | threshold;   ok | ok           ok  ok  threshold;   |
2012       | threshold;   ok | ok           ok  ok  threshold;   |
2013       | threshold;   ok | ok           ok  ok  threshold;   |
2014       | threshold;   ok | ok           ok  ok  threshold;   |
2015       | threshold;   ok | ok           ok  ok  threshold;   |
-----------------------------------------------------------------|

INFO:acro:records:add(): output_0
```

[7]:

survivor	Dead in 2015		Alive in 2015			
grant_type	G	R	G	N	R	R/G
year						
2010	3	47	12	59	24	8
2011	3	47	12	59	24	8
2012	3	47	12	59	24	8
2013	3	47	12	59	24	8
2014	3	47	12	59	24	8
2015	3	47	12	59	24	8

SOCRATES Project – Work Strand 1

- Review of existing SecureLab projects' output requirements
  - Can we pre-identify potential utility of the ACRO tool during project application process?
  - How applicable would the ACRO tool be to the research done within SecureLab?

SOCRATES Project – Work Strand 1

- Phase 1: Review of project applications & administrative documentation
  - Identify what information provided *prior* to project start may suggest possible utility of ACRO
  - Reviewed 25 UKDS SecureLab projects (12 added in Phase 2)
    - Research proposal
    - SRSA-DEA migration form
    - Change of scope request
    - Extension request
    - Approved researcher form

SOCRATES Project – Work Strand 1

- Findings:
  - Most useful information from documentation:
    - Data file type requested
    - Intended software usage
    - Analysis methods
  - Stata was preferred file type
    - Stata preferred software/analysis tool where documented

SOCRATES Project – Work Strand 1

- Phase 2: Review of project output requests
  - Document what types of outputs researchers were requesting
  - Identify utility of ACRO
- 115 outputs across 27 projects
  - 10 Phase 1 projects had no outputs to review (due to project closure)
    - 12 projects added (documentation also reviewed and added to Phase 1 review findings)
  - Up to 5 output requests were reviewed per project
  - 1476 tables and 413 figures

SOCRATES Project – Work Strand 1

- JIRA review
  - Whether output was released
  - How many times request was referred back to researcher
  - Reason(s) for referral
- Output file(s) review
  - Number of output files and page count
  - Total number of tables and figures
  - Types of tables and figures
  - Size of tables
  - Whether statistics were reported in the text
  - Types of statistics in text
  - Likely/actual statistical program used to create output components

Patterns in UKDS SecureLab Output Requests

- Findings
  - Stata most commonly used program (74%)
  - Not uncommon for researchers to submit 15+ tables per output request
    - Maximum: 86 tables in single output
      - Descriptive statistics - 15
      - Summary statistics - 3
      - Interaction terms (p-values) - 3
      - Logistic regression - 43
      - Poisson regression - 19
      - Linear regression - 3
  - Common reasons to reject an output primarily administrative reasons or “missing” information, not statistically disclosive

Patterns in UKDS S

- Findings

Table Type	Total	Number of projects
Frequency tables	138	11
Cross-tab	21	2
Descriptive statistics	181	20
Summary statistics	206	16
Linear regression	152	11
Ordinary least squares	111	8
Survivability table	8	1
Probit regression	2	1
Marginal effects (Probit model)*	28	1
Logistic regression	217	6
Odds ratio*	21	1
Random effects probit regression	1	1
Rotated factor loading	1	1
Tobit regression	75	2
Random effects tobit regression	1	1
VIF model	1	1
Correlation table	21	7
Propensity score matching	3	1
2-stage least squares (incl 1st stage regression)	63	2
Difference-in-difference regression	24	2
Coefficient estimates (unclear which specifically)	15	1
"Parametric" regression (unclear which specifically)	29	1
Poisson regression	29	1
Stratification analysis	6	1
Sensitivity analysis	2	2
Interaction terms (p-values)	3	1
Association estimates	3	1
Robustness checks	2	1
Table of weights	1	1
Zero-inflated negative binomial regression	3	1
Component loadings	5	1
Coarsened exact matching	2	1
Gini coefficient	2	1
Single-group interrupted time series	2	1
Concentration indices	2	1
Dissimilarity indices	1	1
Shapley-Sharrock decomposition of circumstances	1	1
Binary choice generalized linear model with a probit link function	21	1
Poisson pseudo-maximum, likelihood (PPML) Regression	4	1
Pearson's chi-squared	14	1
Factor analysis	18	1
Distance indices	1	1
Random effects Heckman and Mundlak models	1	1
Cobb-Douglas production function	1	1
TFP growth using Haltiwanger methodology model	1	1
Repeated measures ANOVA	15	1
Post-hoc analysis	16	1
Latent growth curve analysis	1	1

Patterns in UKDS SecureLab Output Requests

- Findings

Figure Type	Total
Histogram	2
Survival plot	0
Total	2
Marginal effects	34
Scree plot	5
Kaplan-Meier survival estimates	4
Heat maps	17
Propensity	9
Bar charts	76
Tree cluster diagrams	2
Flow diagrams	10
Predictive margins	6
Pie chart	3
Point plot/scatter plot	46
Line graphs	103
Line graph & scatter plot (combined)	11
Regression coefficients	28
Odds ratio	17
Relative risk ratio	7
Gini coefficient	7
Difference-in-difference estimate	3
Weights	2
Sensitivity analysis	2
Maps	1
Kernal density estimate	4
Structural equation model	14

Patterns in UKDS SecureLab Output Requests

- Findings

Reason for Output Request Rejection	
Presence of below threshold Ns	10
Presence of 0s	1
Missing/incorrect data citation	15
Missing/inadequate descriptions	16
Missing underlying Ns	37
Missing breakdown of Ns	5
Clarification regarding Ns required	15
Missing table/figure/variable labels	8
Presence of minimums/maximums/medians	2
Suppressed values can be recalculated	3
Clarification regarding variables	5
Format	4
Presence of a constant (no longer concern)	1
Presence of embedded documents	1
Missing unit of analysis	1
Presence of UKDS SecureLab server path	1

SOCRATES Project – Next Steps

- Testing the ACRO tool within UKDS SecureLab
- User-testing with current UKDS SecureLab researchers
- Establish what the implementation of ACRO into UKDS workflows would involve

Any questions?

UK Data Service

Thank you.

Allison Tyler allison.tyler@essex.ac.uk
James Rayner jr19302@essex.ac.uk