Planning Research with Administrative Data for Youth Transitions

From Research Priorities to Project Design

Jake Anders

UCL Centre for Education Policy & Equalising Opportunities

2025-04-14

Outline

1. The ADR England Youth Transitions Catalyst
2. The governance landscape
3. Data characteristics and limitations
4. Pre-funding feasibility assessment
5. Development, testing, and outputs
6. Skills and team composition

UCL Centre for Education Policy & Equalising Opportunities

About me

Jake Anders, Professor of Quantitative Social Science at UCL

- Deputy Director, UCL Centre for Education Policy & Equalising Opportunities (CEPEO)
- Co-Investigator, ADR England Youth Transitions Community Catalyst

Research focuses on social inequalities in education and labour market outcomes, using large-scale administrative and survey data.

Contact: jake.anders@ucl.ac.uk

UCL Centre for Education Policy & Equalising Opportunities

Unlocking Youth Transitions Data webinar series

This webinar is part of a series of webinars and events organised by the ADR England Youth Transitions Community Catalyst to support early career researchers in planning and conducting research using administrative data on youth transitions.

Will provide more information on upcoming webinars at the end.

UCL Centre for Education Policy & Equalising Opportunities

The ADR England Youth Transitions Community Catalyst

What is the catalyst?

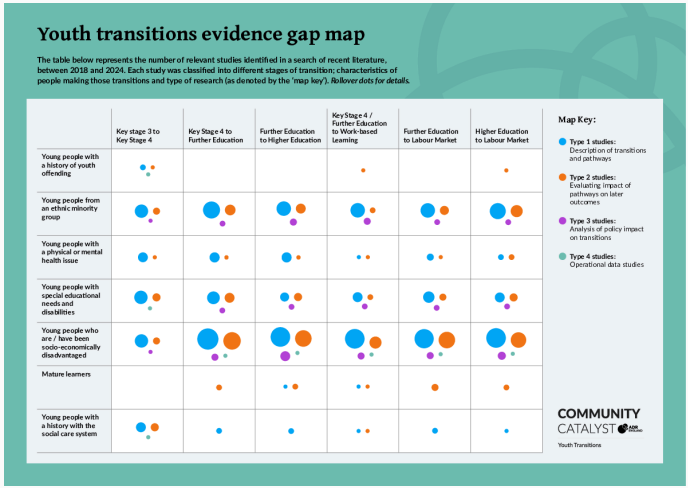
The ADR England Youth Transitions Community Catalyst is funded to build capacity around use of administrative data to explore issues of post–16 education and labour market transitions for young people aged 14–24.

Key aims:

- Create an informed research agenda on youth transitions
- Build capacity for increased use of administrative data in research
- Develop a youth transitions research community
- Address research priorities through primary research

UCL Centre for Education Policy & Equalising Opportunities

Evidence Gap Map



Groups of interest

The Catalyst has identified several groups as particular priorities for research:

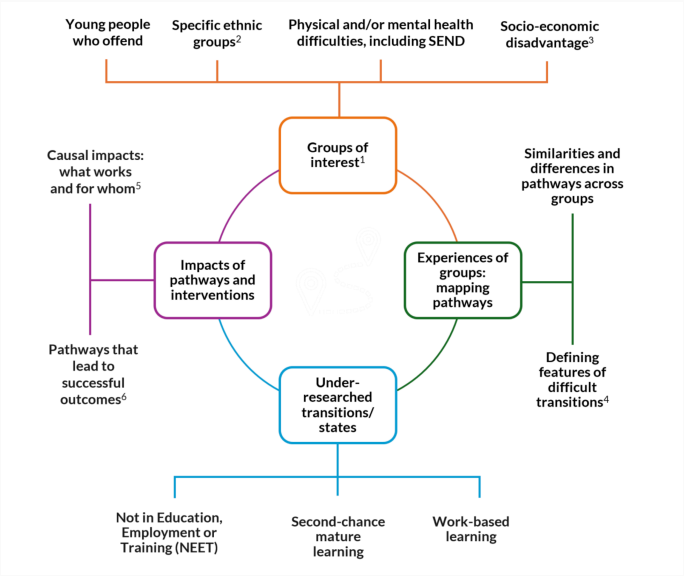
Less research on educational and/or labour market outcomes for:

- mature learners (not technically youth but emerged strongly as priority);
- young people who offend;
- young people with experience of social care or health issues (including physical and mental health);
- some ethnic minority groups.

Lots on socioeconomic status, but something of a disconnect between use of FSM in schools research and multi-dimensional measures of disadvantage later in transitions.

UCL Centre for Education Policy & Equalising Opportunities

Emerging priorities



Emerging priorities

1. Recognise intersectionality and consider individual characteristics and experiences that tend to co-exist or appear to be over-represented in specific transitions.
2. Avoid higher aggregations and consider sub-groups such as Black Caribbean, Black African, Pakistani, Bangladeshi and Gypsy or Irish Traveller and Roma.
3. Consider richer measures of socio-economic disadvantage, e.g., duration of free school meals (FSM) during school instead of a binary FSM eligibility indicator at a fixed point in time.
4. This includes, for example, school exclusion, low attainment, non-continuation in higher education.
5. Any investigation on what works must take into consideration extenuating factors such as location, family and school characteristics. The aim should be to develop a theory of change that identifies the core features of a pathway or intervention (and its context) which can be transferred into similar other settings.
6. What is meant by 'successful outcomes' should be defined at the outset of any research.

UCL Centre for Education Policy & Equalising Opportunities

Especially relevant ADR England Flagship datasets

- Longitudinal Educational Outcomes (LEO) dataset: DfE education records linked to HMRC employment and earnings data
- ECHILD dataset: DfE education records linked to NHS health and social care records
- DfE–MOJ linked dataset: DfE education records linked to Ministry of Justice justice records
- GRADE dataset: DfE/Ofqual/UCAS records

UCL Centre for Education Policy & Equalising Opportunities

Why administrative data?

Administrative data can address research priorities that other data sources cannot but it also brings its own challenges.

What admin data offers:

- **Scale**: entire cohorts, enabling analysis of small groups and rare events
- **Longitudinal coverage**: tracking individuals across time and institutions
- **Linkage**: connecting education, employment, benefits, health, and justice records
- **No attrition**: unlike surveys, administrative records don't depend on ongoing participation

What admin data cannot offer:

- Information not collected for administrative purposes (attitudes, aspirations, parental support)
- Fine-grained measures of mechanisms or processes
- Perfect coverage: some groups are systematically under-represented
- Easy access: governance and practical constraints

UCL Centre for Education Policy & Equalising Opportunities

The governance landscape

Accessing administrative data in England requires navigating a structured governance framework

Understanding this framework will help you design a successful project. Governance requirements shape what is feasible, how long projects take, and what outputs are possible.

Three key elements:

1. The Five Safes framework
2. Trusted Research Environments (TREs)
3. Ethics approval and public benefit requirements

UCL Centre for Education Policy & Equalising Opportunities

The Five Safes framework

Originally developed by the Office for National Statistics (ONS), the Five Safes provide a structured way to manage risk in data access:

Safe element	What it means in practice
Safe people	Researchers are accredited and trained; organisations are vetted
Safe projects	Research has a clear public benefit; purpose is approved
Safe settings	Data is accessed only within a controlled environment (the TRE)
Safe data	Data is de-identified and proportionate to the research need
Safe outputs	Results are checked before release to prevent disclosure of individuals

UCL Centre for Education Policy & Equalising Opportunities

Trusted Research Environments

A TRE is a secure computing environment within which approved researchers can access sensitive data without the data leaving the controlled setting.

Practical implications for researchers:

- You cannot take data away: all analysis happens inside the TRE
- Software available may differ from your usual setup
- Collaboration within the TRE often requires all team members to be separately approved
- Outputs must go through a disclosure control check before you can download and share them outside the TRE
- Internet access is typically absent or heavily restricted inside the TRE

Key TREs for youth transitions research

- **ONS Secure Research Service (SRS)**
- **Secure Anonymised Information Linkage Databank (SAIL)** (Wales)
- **Electronic Data Research and Innovation Service (eDRIS)** (Scotland)
- **UK Data Service Secure Lab** (Mainly survey data, but some admin datasets)
- **UK Longitudinal Linkage Collaboration** (Longitudinal population studies linked to admin data)

Each has its own access procedures, timescales, and software environments.

& Equalising Opportunities

Ethics approval and public benefit

Ethics approval for administrative data research has specific considerations beyond standard human subjects research:

Key considerations:

- Most admin data involves no direct participant contact, but individuals have not consented to research use
- Public benefit must be clearly articulated: underpins the main legal gateways for data access
- Some data (especially health, justice) may require specific statutory gateways
- Consider whose interests the research serves and how findings will be used

Articulating public benefit

Be specific: not "this will help young people" but "this will provide evidence to inform [specific policy/practice decision] for [specific group] by [specific mechanism]."

Because they underpin the legal gateway, public benefit statements are carefully considered.

UCL Centre for Education Policy & Equalising Opportunities

Approval timescales

Plan for data access to take much longer than you expect.

Anecdotal timelines from application to data access:

- **Simple, single-source access** via an established TRE: 3–6 months
- **Linked data** from multiple providers: 6–18 months
- **Novel linkages** or data not previously used for research: potentially longer, or not possible

Timelines are affected by:

- Completeness of your application
- Data provider capacity
- Ethics and governance committee schedules
- Linkage complexity

For grant applications

Build access timelines into your project plan explicitly. Consider contingency plans.

Funders and reviewers familiar with admin data will push back on unrealistic plans!

If it involves access to data outside a well-trodden path, data and funding applications will likely be strengthened engaging with the data provider(s) (and providing evidence of this engagement).

UCL Centre for Education Policy & Equalising Opportunities

Application processes for SRS: where?

- Applications are managed by ONS with applications via the Project Accreditation Service for SRS (PASS)
- There is helpful project application guidance and an exemplar project application to help researchers understand the level of detail needed
- There is a version of the application form as a Word document which you might find useful for collaborative preparation of an application, before pasting the responses into the online system

Different for entirely DfE-owned data (e.g., NPD) where applications continue to be managed by DfE.

UCL Centre for Education Policy & Equalising Opportunities

Application processes for SRS: what?

Information you will need to provide:

- Detail about your project – RQs addressed, methods used, etc. – to enable another researcher to understand plans and judge basic 'value'
- Evidence of ethical consideration of your project
- Evidence that your research will be acceptable to the public
- Details of how you anticipate your project might have public benefit
- Project biases and limitations
- Potential harm and risks of your project affecting individual's protected characteristics
- Also need to complete a variable request form (VRF), specifying data tables, variables and years of data needed
- Details of any additional datasets you would need to ingest need to be attached on a separate external data request form (although for many datasets individual-level linking is not allowed)

Applications are initially reviewed by ONS staff to assess feasibility and public good before being passed to data owners for approval and then to Research Accreditation Panel for final sign-off

UCL Centre for Education Policy & Equalising Opportunities

Data characteristics and limitations

Administrative data was collected for administrative purposes — not to answer research questions.

This has fundamental implications for:

- What is recorded and how
- Who appears in the data
- How consistent records are over time
- What linkages are possible

Understanding these characteristics is important to planning a feasible project.

UCL Centre for Education Policy & Equalising Opportunities

Administrative origin and its consequences

Because admin data is collected for administrative purposes, definitions and coverage reflect administrative rather than scientific value:

What this means:

- Variables measure administrative categories, which may not map cleanly onto scientific constructs
- Coverage reflects who interacts with a system, not who exists in a population
- Quality varies by how important accurate recording is for administrative purposes
- Changes to systems, policies, or recording practices can create apparent trends that are artefacts

Example: NEET

NEET status is not always well-identified by educational institutions' records. NCCIS data has improved identification of activities of groups outside this. However, its quality is highly variable across local authorities.

See this 'Data Explained' on 'Identifying and studying young people who are Not in Education, Employment or Training (NEET)' for more discussion.

UCL Centre for Education Policy & Equalising Opportunities

Sample selection in administrative data

Who appears in administrative data — and who doesn't — is not typically random:

Common selection issues in youth transitions data:

- School/college records only cover those enrolled — early leavers may disappear
- Apprenticeship records depend on employer registration — informal work-based routes are not captured
- Benefits data covers only those who claim — eligible non-claimers are absent
- Health records reflect service use, not health need
- Justice data captures those who are caught and processed — not all offending behaviour

The groups most at risk of being missed

Often the most disadvantaged young people are also least likely to appear consistently in administrative records. Emphasises that we should typically not assume that simply dropping those with missing data will be an unbiased way to proceed.

UCL Centre for Education Policy & Equalising Opportunities

Missingness and temporal change

Scoping temporal coverage

> A common issue: researchers design a study requiring ten years of linked data, only to discover that the relevant records were only linked from a certain year onwards, or that a key variable only exists from a certain year onwards.

Missingness in admin data:

- Missing values often reflect administrative processes, not randomness — a missing SEND flag may mean no assessment was done, not that no need exists
- Patterns of missingness may differ across geographies, institutions, and time
- Do not assume missingness is ignorable — it usually isn't

Longitudinal complications:

- Recording practices, variable definitions, and systems change over time
- What looks like a trend may be a system change
- Linkage keys (e.g., URNs) may be inconsistent across time

UCL Centre for Education Policy & Equalising Opportunities

Useful resources

ADR UK have published a collection of ‘data explained’ outputs, describing features such as quality issues, changes over time, and linkage caveats.

Examples:

- Understanding post-16 learning spells and qualification aims to estimate returns to learning using LEO
- Exploring the dynamics of school absenteeism and crime using NPD-MOJ linked dataset

Really useful to review any that may be relevant to your plans.

UCL Centre for Education Policy & Equalising Opportunities

Pre-project feasibility assessment

A key challenge: you cannot see the data while planning a project — but you need to design your project around it.

Strategies exist for assessing feasibility before access.

Motivating question

Can you estimate, before access, whether the key subgroups in your study will be large enough to support the analysis you want to do?

If your research question requires detecting a 0.1 SD effect in a subgroup that makes up 2% of the population then very important to work out if the data plausibly contains enough observations.

UCL Centre for Education Policy & Equalising Opportunities

Feasibility without data access

Use documentation:

- Data dictionaries and variable lists from data providers
- ‘Data explained’ publications
- Especially check what years are available and which variables are in which years

Use aggregate statistics:

- Published statistics can help you estimate sample sizes and subgroup frequencies
- Will your group of interest be large enough to support your intended analysis?

Use published research:

- What have others been able to do with this data? What constraints did they encounter?
- Pay attention to methods sections and data appendices — researchers should document limitations

UCL Centre for Education Policy & Equalising Opportunities

Researcher networks and data owner contacts

You don't have to figure everything out from documentation alone:

Talk to other researchers:

- Colleagues who have used the same data are invaluable sources of practical knowledge
- ADR UK communities of practice and researcher networks exist specifically for this
- ADR Youth Trainings Community Catalyst events and community

Engage with data providers early:

- Data providers will generally be pleased to address questions you have (although you might have to chase)
- This early engagement can clarify feasibility to avoid wasted time on an unfeasible project
- As noted above, can also be useful to flag this in funding applications

UCL Centre for Education Policy & Equalising Opportunities

Development and testing strategies

Even once access is approved, developing and testing analytical code inside a TRE is slow and effortful. Good preparation before access significantly reduces wasted time inside the environment.

Plan around TRE constraints

Think in advance about:

- What software is available inside the TRE?
- Can you use your preferred packages or will you need to adapt?
- Do you need to adapt sequencing of work given need to clear outputs?

UCL Centre for Education Policy & Equalising Opportunities

Synthetic data and code preparation

Synthetic data:

- Data providers are increasingly working to offer synthetic versions of their data — structurally similar to the real data but containing no real individuals
- Use synthetic data to develop and test your planned data cleaning and analysis code before access
- Allows you to identify and debug some issues early: much faster to debug outside a TRE
- Won't catch everything!

Code sharing:

- Search for existing code from other researchers using the same data, although this can be scant outside TRE
- SQL code and guide to prepare LEO to study post-16 activities from FFT Education Datalab
- Code to help make LEO 'research ready' available in SRS Code Library
- Reproducible research examples give you a head start and help you avoid known pitfalls

UCL Centre for Education Policy & Equalising Opportunities

Outputs and dissemination constraints

All outputs from a TRE must go through **statistical disclosure control (SDC)** checks.

What SDC typically involves:

- Suppression of small cell counts (typically $n < 5$ or $n < 10$, depending on provider)
- Rounding of counts and percentages
- Review of regression outputs (residual degrees of freedom)
- All apply to graphical outputs as well as tables (will need tables of counts underpinning the graphs)

Practical implications:

- Analyses of small subgroups may produce many suppressed cells
- Plan your table and figure structure with SDC in mind
- Budget time for output checking

Acknowledgement requirements

Data providers require specific acknowledgement language in all publications using their data. Check requirements for each dataset used and for the TRE itself. Non-compliance is a breach of the terms of access so can theoretically jeopardise future access. On a more public spirited note, it is also important for understanding use of data and for demonstrating the value of data access to funders and the public.

UCL Centre for Education Policy & Equalising Opportunities

Planning for output checking

Build output checking into your project timeline:

- Allocate dedicated time in your project plan for each round of output checking
- Consider submitting outputs in batches: one large well-prepared submission is usually faster than multiple small ones (but avoid something too big, which they may wish to break up)
- Prepare a clear output checking request

Consider outputs early:

- The best time to think about SDC requirements is when designing your analysis, not when writing up
- If a finding depends on a cell count that will be suppressed, you may need to redesign the analysis

Output types that typically require extra attention:

- Cross-tabulations with many cells
- Any analysis focusing on very small groups
- Linked-data analyses where multiple suppression rules might apply

UCL Centre for Education Policy & Equalising Opportunities

Skills and team composition

Working effectively with administrative data in a TRE environment requires some specific skills. Important to consider explicitly when building a project team.

What skills matter?

Technical skills:

- **Data management**: admin datasets are large and often messy; strong data cleaning skills are essential
- **SQL**: some TREs and data sources require or strongly benefit from SQL knowledge for querying large databases
- **R or Python**: have been moves to prioritise provision and availability of *open source* software in TREs

Domain and methodological skills (above and beyond those needed with non-admin data):

- Understanding of the administrative systems that generated the data
- Experience with disclosure control processes

Training:

- ADR offer training and development opportunities
- Consider whether your team has gaps and build in training time

UCL Centre for Education Policy & Equalising Opportunities

Summary

Working with administrative data for youth transitions research can be highly valuable to address important research priorities. However, it requires specific planning and preparation.

Key takeaways:

1. **Research priorities**: YT community catalyst has identified clear gaps that admin data is well-placed to address
2. **Governance**: understand Five Safes, governance and TRE workflows and their relevance for your project
3. **Data characteristics**: admin data has limitations that must be understood and addressed, not assumed away
4. **Feasibility**: assess feasibility using documentation, aggregate statistics, and researcher networks before committing
5. **Planning**: build in time for approvals, synthetic data testing, and output checking; they take longer than expected
6. **Team and skills**: be explicit about what skills your project needs

UCL Centre for Education Policy & Equalising Opportunities

Some Resources

- ADR UK
- ADR UK Data Catalogue
- ADR ‘data explained’ publications
- ADR England Youth Transitions Community Catalyst
- ADR YT CC: Data Navigation Tool
- ADR YT CC: Knowledge Hub
- ADR YT CC: LinkedIn group
- ONS Secure Research Service

UCL Centre for Education Policy & Equalising Opportunities

Upcoming events

Unlocking Youth Transitions Data webinar series

- **TOMORROW**: Using the PICO (Population, intervention, comparator and outcomes) framework in LEO studies of post-16 learning
- Next week: Creating and using proxy measures of socioeconomic background in LEO

UCL CEPEO training events

- Monday 20th April: Practical introduction to LEO using R
- Upcoming training on NPD data (TBA)
- Upcoming training on using administrative data for evaluation (TBA)

UCL Centre for Education Policy & Equalising Opportunities

Questions

Thank you

Jake Anders — jake.anders@ucl.ac.uk

Feel free to ask about anything covered today — or anything related to planning youth transitions research using administrative data.

Stay in touch

Details of future webinars and events will be circulated via the ADR UK networks.

For questions after today, email is the best way to reach me.

UCL Centre for Education Policy & Equalising Opportunities