Quality issues related to linkage of administrative data

Purpose

This paper aims to raise awareness of the impacts of not understanding the quality of linkages involving administrative data, set out recommended work to address these issues.

The future vision for ONS statistics involves the integration of administrative data with survey and other data sources to provide the highest quality outputs available to users on a more frequent and granular basis than is currently possible from survey data alone. In the context of utilising additional administrative data, the Statistical Infrastructure and Design hub within Methods and Quality Directorate (MQD) in ONS has identified a need to prioritise the fundamental building blocks of this system: the quality of linkages involving administrative data. A strategy for managing user expectations around the quality of administrative data quality is needed that reflects consideration and communication of quality at all stages of linkage from design through to analysis. The paper recommends a number of specific areas of research and development to improve the quality of administrative data linkage which has been endorsed by boards internal to ONS. This work is highly relevant to the National Statisticians 2023 Recommendation and so we invite MARP to review this paper for sight of additional ONS work in this space. Sharing this work with MARP also ensures that via transparent publication onto the UKSA website, this work can be accessed and referenced to in the 2023 consultation document.

Recommendations

The paper highlights that administrative data linkage quality has limits that may never be fully resolved, so work needs to be done to manage user expectations and to ensure that quality is considered at all stages of linkage from design through to analysis.

It is recommended that further work is needed in the following areas:

- Improvement of linkage methods for administrative data
- Further development of the ONS clerical matching service to facilitate high quality linkage and estimation of quality metrics in a range of scenarios
- Improvement of data quality information, including precision and recall metrics, for administrative linkages
- Inclusivity of linkage
- Work to investigate quality risks and issues around using common IDs derived from linkage of separate datasets to the Reference Data Management Framework (RDMF) to join datasets
- Development of supporting materials to enable analysts to better articulate both their linkage quality requirements as well as impacts of linkage on their analysis
- Investment in improving the quality of our main data sources to facilitate high quality linkage
- Discuss whether a bid should be made to PIC to fund this work in the absence of specific programme funding.

Background

Quality assessment of administrative and alternative data at the source-level is well-established in line with best practice across ONS. Issues are reported in the context of the core data quality dimensions established by the European Statistical System and Data Management Association UK (DAMA(UK). As such, biases are generally well-understood, although not always quantified, e.g., Patient Register: quality assurance of administrative data used in population statistics, Dec 2016 - Office for National Statistics; Higher Education Statistics Agency data: quality assurance of administrative data used in population statistics, Feb 2017 - Office for National Statistics.

However, due to the absence of any single, comprehensive, alternative data source to replace the rich traditional data sources like the census, linkage of alternative data will inevitably play a foundational role in the transformed statistical system. For example, the Dynamic Population Model explicitly assumes that alternative sources (admin, surveys) can be linked in order to facilitate estimates of population size (stocks) through compiling Statistical Population Datasets (SPDs), with the use of surveys to estimate bias in the SPDs.

This means that it is not enough to simply understand source data quality when considering the use of administrative data.

Unfortunately, the quality implications specific to linkage of administrative data are far less well-understood e.g., the process of linking two administrative sources may create additional quality issues which were not present in the source datasets, especially if the linkage process is poorly designed.

This is important because, in line with Government Data Quality Framework principles, data should be managed across its lifecycle and users should assess data quality at every stage,

take proactive measures to improve quality when issues arise, and adopt appropriate assessment measures at each stage rather than applying a one-size-fits-all approach to quality assurance. The aim is to be able to provide clear data quality information in relation to the linkage and describe its impact on use of the data. We do not feel that this is currently developed to an extent to enable analysts to make such an assessment and know what to do as a result.

It should be noted that linkage is not the only process that needs to be considered when understanding the quality of a linked ("composite" or "integrated") dataset because there are many other sources of error, but the aim of this paper is to highlight concerns specific to linkage.

Linkage of alternative data sources is achieved using both traditional and non-traditional methods and may be deterministic/rule-based or probabilistic but, crucially, Methodology and Quality Directorate have identified that the methods for assessment of linkage quality currently used for survey data cannot be applied. The clerical resolution methods which underpin Census and survey linkage are not as accurate when working with administrative data sources (Annex A provides an illustrative example). Ultimately, this causes ambiguity in linkage accuracy and bias where we are unable to confidently identify errors as well as hindering our ability to link administrative data in the first place (Annex B presents a comparison between metrics from a traditional linkage and one involving administrative data). We feel this has not been fully appreciated in the design of the transformed statistical system, with implications for any proposed alternatives to the traditional census, and therefore needs to be urgently addressed.

Discussion

Broadly, these are the known areas where linkage of administrative data sources is occurring within ONS:

- Combination of sources to create indexes as part of the Reference Data Management Framework (RDMF), e.g., the creation of the Demographic Index (DI) by combining Patient Demographic Service, School Census, Higher Education Statistics Agency etc. data, which would ultimately be used as a basis for SPDs, amongst other uses.
- Linkage of alternative data sources with the RDMF (via RDMF matching services or other standardised linkage methods) to attach a common ID to allow joins with other datasets. This is the proposed model for creation of data products for the Integrated Data Service (IDS).
- Creation of bespoke linked datasets by linking various data sources to allow analysts to answer specific research questions e.g., the Public Health Data Asset, Education-Health Asset etc.
- Linkages not covered above e.g. where alternative data is used to supplement/replace survey sources for a specific purpose e.g., linkage of the Valuation Office Agency (VOA) number of rooms variable to the 2021 England and Wales Census data to replace the census number of rooms question.

We are currently unable to confidently report the quality of linked datasets produced by these means. This is because the clerical resolution methods that underpin census and survey linkage are not as accurate when working with administrative data. Clerical resolution

is used to estimate key quality metrics for data linkage like precision and recall[1]. It is possible to calculate precision and recall for administrative data linkages but, due to the lack of corroborating evidence in administrative data to determine whether a match should really be a match (e.g., marital status, alternative address), there is uncertainty in any reported precision and recall figures. This contrasts with census and survey linkage where we can more confidently report precision and recall. Furthermore, it is currently unclear how to produce quality metrics when joining datasets via an already linked asset e.g. linkage achieved via the RDMF, which further adds to the uncertainty around quality when using integrated datasets derived by these means.

Being unable to confidently report the quality of a linked dataset means we will be unable to inform our users about potential bias and error in the datasets. This could lead to users making conclusions based on the data that are inaccurate, misleading, or discriminatory. Publishing statistics like these will eventually harm our reputation and position as a world-leading institute in statistics and lead to poor research outcomes and policy decisions.

We already know that linkage error disproportionately affects certain groups (linkage bias). Based on observations from previous linkages (e.g. PDS to 2021 Census linkage), we know that Asian, Mixed, other and Black ethnic groups are harder to link in administrative data than the White group, leading to datasets that exclude or incorrectly link them. This could have wide ranging impacts because Data Linkage is frequently used to inform health outcome policy. For example, the Asian ethnic group also has a high prevalence of diabetes. If we created a linked dataset to investigate this prevalence using administrative data, we are likely to end up with a dataset that excludes a proportion of the Asian ethnic group – which would suggest that diabetes is a smaller problem than it is. Since quality assessment is limited in administrative data, we would currently be limited in our ability to test whether our estimates on diabetes prevalence are accurate or measure our confidence in them. We would be publishing statistics without being able to estimate their quality or confidence.

Basing policy decisions on this diabetes dataset may lead to changes in interventions and support, which could have negative impacts on health outcomes and ONS/government reputation. There is also a risk that this could introduce issues in complying with the Equality Act 2010.

There is currently some work underway to develop methods and increase the knowledge base around administrative data linkage:

- DI to Census/Census Coverage Survey (CCS) linkage exercise – while being the first step to development of a generalised DI matching service, this is also the first time that ONS has attempted to estimate uncertainty in precision and recall figures derived from linkage of administrative data. This will be a best-case scenario because the exercise involves census data, which contains corroborating information that can be used in clerical resolution, so further work is needed to build on this.

[1] Precision is the proportion of assigned links that are true matches. Recall is the proportion of true matches that are correctly identified as links.

- Development of a generalised DI linkage method - this will build on the DI to Census/CCS linkage and ensure that DI linkage methods are implementing best practice.
- Proof of concept for Health Integrated Data Asset (IDA), where we are trying to understand the quality, coverage and practical issues involved with creating IDAs via the Demographic Index
- Research into Machine Learning methods for Linkage - aim to reduce the requirement for clerical matching by using machine learning to achieve high quality automated matching. Potentially applicable to all linkages.
- Longitudinal Labour Market Survey (LMS) linkage evaluation – linkage of waves of LMS independently and comparison with linkage via the DI. This piece of work should give us some good evidence about the quality of the existing linkage method for datasets to DI.
- Development of approaches to enable more accessible ways of evaluating linkage e.g. Bias analysis tool, quality toolkit.
- Development of an over-arching linkage strategy.

This work, apart from the development of an over-arching linkage strategy, has been commissioned by specific business areas within ONS so does not necessarily consider the wider picture with a long-term view. We therefore feel that further work is required in the following areas to ensure that linkage of administrative data is considered in a holistic way and that proper foundations are in place to meet future linkage requirements across government, including via IDS.

Work required:

- Further research into improving linkage methods for working with administrative data – this will maximise the quality of linkage and mitigate against the impact of linkage biases with a focus on the whole linkage space rather than simply the DI, so will have benefits for bespoke linkages, linked data assets and any other applications where RDMF linkage may not be appropriate.
- Further development of the ONS clerical matching service to provide a resource to facilitate high quality linkage and estimation of quality metrics in a range of scenarios
- Improve data quality information
- Further empirical calculation of the uncertainty in precision and recall from administrative data linkages, including those that do not involve census/survey data – this will help us to understand the magnitude and impact of this issue and inform the methods being developed to improve linkage as well as assess linkage quality
- Further work to build methods for identifying issues with administrative datasets such as bias or accuracy and design surveys to capture this missing information – this will give us more certainty about the quality metrics produced for administrative data linkage, which will allow users to explain the impact of linkage on the use of the data in line with Government Data Quality Framework
- Inclusivity in linkage work programme – set up series of work packages to focus specifically on inclusivity issues e.g. development of methods for linkage of “hard to link” groups, analysis of biases affecting groups with protected characteristics etc. This will ensure that we are considering linkage in the context of the Equality Act and taking specific action to ensure inclusivity of statistics based on linked data.

- Work to investigate quality risks and issues around using common IDs derived from linkage of separate datasets to, for example, the RDMF to join said datasets e.g. how would you estimate and report quality from linkages achieved in this way?
- Improve communication of quality related to administrative data linkages, including upskilling of linkage experts to understand and articulate quality issues, and development of accessible supporting materials in conjunction with users e.g. guidance for analysts using linked administrative data. This will enable analysts to better articulate both their linkage quality requirements as well as impacts of linkage on their analysis.
- Invest in improving the quality of our main data sources by, for example, working with data administrators and suppliers to maximise the variables that are acquired for linkage to increase the chance of achieving high quality linkages and reduce uncertainty in quality metrics. Work could also include assuring inputted data quality and supporting development of better infrastructure to allow better data entry. This will maximise the quality of linkage, support quality assessment of linkage, and mitigate against the impact of linkage biases.

List of Annexes

Annex A Illustrative example of the difference between clerical resolution in census data compared with administrative data

Annex B Quality metrics for Census-CCS linkage compared with Demographic index to Census/CCS

Annex A - Illustrative example of the difference between clerical resolution in census data compared with administrative data

In this situation, we are trying to decide if a person with a common name and some error in their data links to another record. This example, while synthetic for disclosure purposes, is representative of many of the uncertain links that we use clerical matching to resolve and accurately represents the data fields available in different datasets quoted.

Example: A matching person going through a divorce

Using the Census and CCS data, clerical matchers can confidently link the individual using supplementary information – despite errors in the data, we can clearly see that this is the same person, but they have gone through a divorce.

Census and CCS data

	Census	CCS
Forename	Johnathan	John
Middle name		
Surname	Smith-Jones	Smith
Date of Birth	1st April 1992	1st April 1991
Sex	Male	Male
Full Address	**3 High Street, Norton, West Land, WL3 6GH**	Flat 5, 3 North Road, West Land, WL5 13FG
Alternative Address		
Address 1 Year Ago	3 High Street, Norton, West Land, WL3 6GH	**3 High Street, Norton, West Land, WL3 6GH**
Marital Status	**Married**	**Divorced**

Country of Birth	England	England
Ethnicity	White European	White European
Employment	Employed	Student
Industry	Catering	Student
Occupation	Chef	Student
Other household members	Mary Smith, 37 **Fred Smith, 10** Amy Smith, 6	**Fred Smith, 10**

However, in the administrative datasets the decision is much less certain – John Smith is a common name so how do we know there aren't two with the same birthday living in 'West Land'? Since his divorce, his surname has changed, but because marital status information is not captured in the administrative sources, we have no way of knowing about this change of status. We are therefore unable to confidently say that these 3 records belong to the same person and would be unable to confidently say if we have made a mistake for our quality assessments.

Administrative Dataset made by combining 4 in the clear datasets used for population estimation in PMST

	Higher Education Stats Authority Dataset	Patient Demographics Service Dataset	English School Census	Welsh School Census
Forename/ First names	John	Johnathan	Johnathan	
Middle name	Not asked			
Surname	Smith	Smith-Jones	Smith	
Date of Birth	1st April 1992	1st April 1991	1st April 1992	
Gender	Male			
Sex		Male	Male	
Postcode	WL5 13FG			
Full Address	Not asked	3 High Street, Norton, West Land, WL3 6GH	3, 5 North Road, West Land, WL5 13FG	
Other postcode	WL5 13FG			

Annex B - Quality metrics for 2021 Census-CCS linkage compared with Demographic index to Census/CCS

These results compare quality metrics derived from a traditional linkage exercise (Census to CCS) to those from linking Census/CCS data to administrative data. Note that differences are likely to be even more pronounced where linkage involves two alternative/administrative sources because the presence of census data in the second scenario mitigates somewhat for issues with alternative data.

Note that precision is a measure of how many of the links that were made are correct. Recall is a measure of how many of the possible correct links were made.

It is also important to note that estimates of precision and recall are often very high so margins appear small, but even slight deviations will impact on whether quality targets are met. This is reflected in the fact that quality targets are often defined to two decimal places e.g. for 2021 Census to CCS matching, targets were set at 99.99% and 99.75% for precision and recall respectively.

Uncertainty, or where the decision to declare two records as a match is subjective, has been factored into all precision and recall estimates, but is more apparent in the clerical matching stage.

The results show that there is higher uncertainty in the precision and recall figures estimated for the DI-Census/CCS linkage compared with the Census-CCS linkage due to the difficulty in clerically reviewing administrative data linkages as illustrated in Annex A.

2021 Census to CCS linkage quality metrics

Matching method	Precision	Lower bound of 95% CI	Difference (p.p.)
Automatic	99.995%	99.968%	0.027
Clerical	99.352%	98.874%	0.478
Overall	99.963%	99.913%	0.05

Matching method	Recall	Lower bound of 95% CI	Difference (p.p.)
Overall	99.959%	99.928%	0.031

2021 Census/CCS to Demographic Index linkage quality metrics

Note that automatic matching is assumed to have 100% precision because during method development any match-key that let through an incorrect link was removed or tightened and rechecked.

Metric	Best Case Estimate	Worst Case Estimate	Difference (p.p.)
Precision of Automatic Links	100%	100%	0
Precision of Clerical Links	95.6%	92.2%	3.4

Metric	Best Case Estimate	Worst Case Estimate	Difference (p.p.)
Recall of Automatic Links	91.9%	91.6%	0.3
Recall of Clerical Links	96%	89.5%	6.5

Overall figures for 2021 Census/CCS to DI linkage

Metric	Best Case Estimate	Worst Case Estimate	Difference (p.p.)
Precision	99.7%	99.4%	0.3
Recall	99.7%	99.1%	0.6