Evaluating Statistical Quality in the Demographic Index

Demographic Index v2.0 and Methodology Division Quality Assessment

Strategic Statistical Design team: Rosalind Archer, Elise Kenning, Siân Lloyd, Pratibha Vellanki

DI team: Paul Groom, Santosh Ankeru, Jack Gibson, Alex Key, Tomos Lewis, Kostas Loukas

Expert Panel: Louisa Blackwell, Fern Leather, Nicky Rogers, Rachel Shipsey

Oct 2022: This paper has been updated in the light of feedback from MaRAG and other colleagues at ONS

Key Messages of Paper

Purpose: This paper presents recommendations and proposed quality metrics from the first iteration of the Methodology Division Quality Assessment (MDQA) for the Demographic Index (DI). The overall purpose of this work is to ensure that the DI can be used to make high quality statistics. This requires that the statistical quality of the DI is known, and that users can understand and report on the quality of what they make when they use the DI.

Recommendation: The work to establish the quality of the DI, and to optimise it, is substantial. The scope of the first iteration is to review the design of DI v2.0, and recommends the following next steps:

1. Establish quantitative measures for quality in the DI
2. Test the DI design
3. Establish governance
4. Support users to use the DI appropriately

Key Asks of MARP: we wish MARP to review our recommendations and proposed next steps. We are keen to ensure that this work is heading in the right direction such that the DI can be used effectively in ONS.

Relevance to National Statistician 2023 Recommendation: ONS will provide evidence to the National Statistician to inform his recommendation for whether a Census will be conducted in 2031. This paper is to be included in that evidence because it relates to the quality of the DI, and the DI underpins the proposed method for making population estimates in the absence of a Census.

Specifically, the DI is used to create a Statistical Population Dataset[1], which in turn is fed into a Dynamic Population Model[2], which produces population estimates. Therefore, the quality of the DI will feed directly into the quality of the SPDs. For instance, if the DI fails to link individuals across sources, resulting in two entries on the DI for a single real person, then that may feed into the SPDs depending on the rules used to construct the SPD.

Acknowledgments: The authors wish to thank all those who have helped us to develop this work. In particular, our expert panel: Louisa Blackwell, Fern Leather, Nicky Rogers, and Rachel Shipsey.

[1] Statistical Population Dataset methodology has previously been taken to the Methodological Research Assurance Panel (MARP). Word version available on request.

[2] Dynamic Population Model methodology has also been to MARP. Word version available on request.

Executive Summary

Context - Background: The Demographic Index (DI) is a composite data source, built from a range of admin sources using cuts of data from approximately 2016-2021[3]. It provides a solution for analysts who wish to work with linked data, but who – for disclosure reasons – may not have access to Personal Identifiable Information (PII), which is typically used to link records. It is a high-profile piece of work being undertaken for the Integrated Data Service (IDS), and there is a strong appetite for it within ONS.

From a methodological point of view, we need the DI to support research into – and the production of – statistics that have known statistical quality. That is, statistics that have quantitative measures of error and uncertainty, such as bias and variation. The DI must also support multiple statistical uses, such as point-in-time analysis and longitudinal analysis.

A Methodology Division Quality Assessment (MDQA) was begun for the DI in February 2021, with a view to assessing the quality of the DI and its fitness for purpose according to our methodological needs. The MDQA for the DI is led by the Methodology and Quality Directorate (MQD), in close collaboration with a team of experts, and with the team who have built the DI.

Context – the challenge: To assess the use of the DI to produce statistics and to support research, it is necessary to understand the statistical quality of the DI itself. However, methods for measuring its quality have not been established.

Our starting point was to fully describe and examine the design of the DI (v2.0), and thus identify where issues of quality might arise and how the DI quality might be measured. This work has resulted in the content of this paper, which constitutes the first iteration of the MDQA.

Overview of paper: In this paper, we will:

- describe the DI v2.0 and its design (data, high level build, stages in process)
- raise key observations relating to quality, and specific recommendations that address these points
- summarise recommendations into five major themes (see below)
- describe key challenges and how recommendations may address them
- conclude with recommendations on how we should proceed

Our recommendations fall into the following major themes:

1) measure statistical quality – we have identified metrics that we believe could be developed to measure the DI quality (see Annex A1) and reported to users on each update; we recommend that they be obtained and validated using data where the DI has been linked to a high-quality data source
2) testing of design by simulation – since the build of the DI is complicated, there are multiple instances where it is difficult to predict the impact of a design

[3] 2016-2020 is the window over which all sources are available. See Figure 1 for more details.

decision on the DI; we recommend that existing tools be used to simulate and test the DI design (see Testing, below)

3) governance – the DI is still relatively young, and is developing all the time; we recommend that governance be established to ensure that development is transparent and that work to understand and improve quality keeps pace
4) research – we expect that key pieces of novel research are required to allow the DI to properly develop, both in scale and in improved statistical quality; we especially recommend research into graph databases and longitudinal analysis
5) usage guidance - there is a pressing need for users to understand the DI, and become adept in using it; we recommend work to support users, especially in terms of documentation and development of use cases

Impact of work in ONS: The quality of the DI affects any work to support the transformation of population statistics in ONS, and analysts who wish to use the DI to conduct research or produce population statistics as part of a transformed system.

Key related pieces of work include:

- the Reference Data Management Framework (RDMF), and indexes within it – which will form a major part of the Integrate Service for the Integrated Data Platform
- Statistical Population Datasets (SPDs) – as mentioned above
- work to link the DI to Census and CCS (DI-CC), for the purposes of evaluating SPD and the DI quality
- proof of concept work for the Census Data Asset, in the form of the Refugee Integration Outcomes (RIO) cohort study

Previous papers/related work:

These are available on request, and include:

- a metrics supplement, containing initial descriptive metrics for the DI – supplied to MARP with this paper
- the release note for the DI v2.1, which also includes metrics – also to be supplied to MARP
- interim papers from MDQA:
    - Paper for Longitudinal Scientific Advisory Panel
    - Comprehensive paper written to provide information to internal expert panel, including full details for the DI v2.0 design and all recommendations
- paper describing linkage of the DI to Census and CCS (DI-CC)

Main Paper

Introduction: This paper covers the first iteration of the MDQA for Demographic Index (DI). The purpose of the MDQA is to assess the quality of the DI – the data, the design, and its fitness for use. For the first iteration, we have reviewed the design of the DI v2.0.

Caveats and limitations: Our work to date covers the process by which unencrypted data are built into a final "DI linktable"[4]. It does not fully cover "hashed" [5] data in the DI, and it does not provide all the information that users will require in order to use the DI.

The method used for hashed linkage in the DI[6] has been reviewed, but the quality of hashed linkage in the DI is not yet known. Work to assess the quality of hashed linkage in the DI is currently beyond scope, due to reasons of resource and limits in our data sharing agreement with DWP. We return to the impact of hashed data on quality later, in "Challenges".

A thorough description of how to use the DI is beyond scope because this will vary depending on the application; instead, at this stage we seek to understand the quality of the DI itself. In the future we anticipate that the natural progression of work will be to understand how the DI quality feeds into outputs. While this work is currently out of scope, we do offer some "Guidance on Usage", later.

The design – the data: the DI v2.0 is a patchwork of linked data that spans approximately 5 years, with one year per "cut" of data, and including the following sources:

1. Personal Demographic Service (PDS): National Health Service
2. Higher Education Statistics Agency (HESA): Student enrolments for tertiary education – university students
3. English School Census (ESC), and Welsh School Census (WSC) – school students not in private education
4. Client Information System (CIS[7]): data from the Department of Work and Pensions, covering Pay As You Earn and benefits data
5. The Births Register
6. The Individualised Learner Record (ILR)

CIS data are supplied hashed, and will be discussed later.

The current order for building the DI v2.0 is shown in Figure 1. PDS 2016 is used as the initial starting dataset for the index because PDS captures many people in the population, meaning that as subsequent data are added there is a good chance that a link can be found for records. The intention for the ongoing build of the DI is to add data as it becomes available ("rolling" the DI).

[4] This is the end product of the DI build, but will not be the same as the data that a user receives, since part of the rationale for the DI is that users receive sufficient information for their analysis needs and no more than that.
[5] That is, with Personally Identifiable Information (PII) and source ID encrypted
[6] Shipsey, R. & Plachta, J. (Updated 16 July 2021) "Linking with anonymised data – how not to make a hash of it" (Link)
[7] Not to be confused with the Covid Infection Survey, also "CIS"

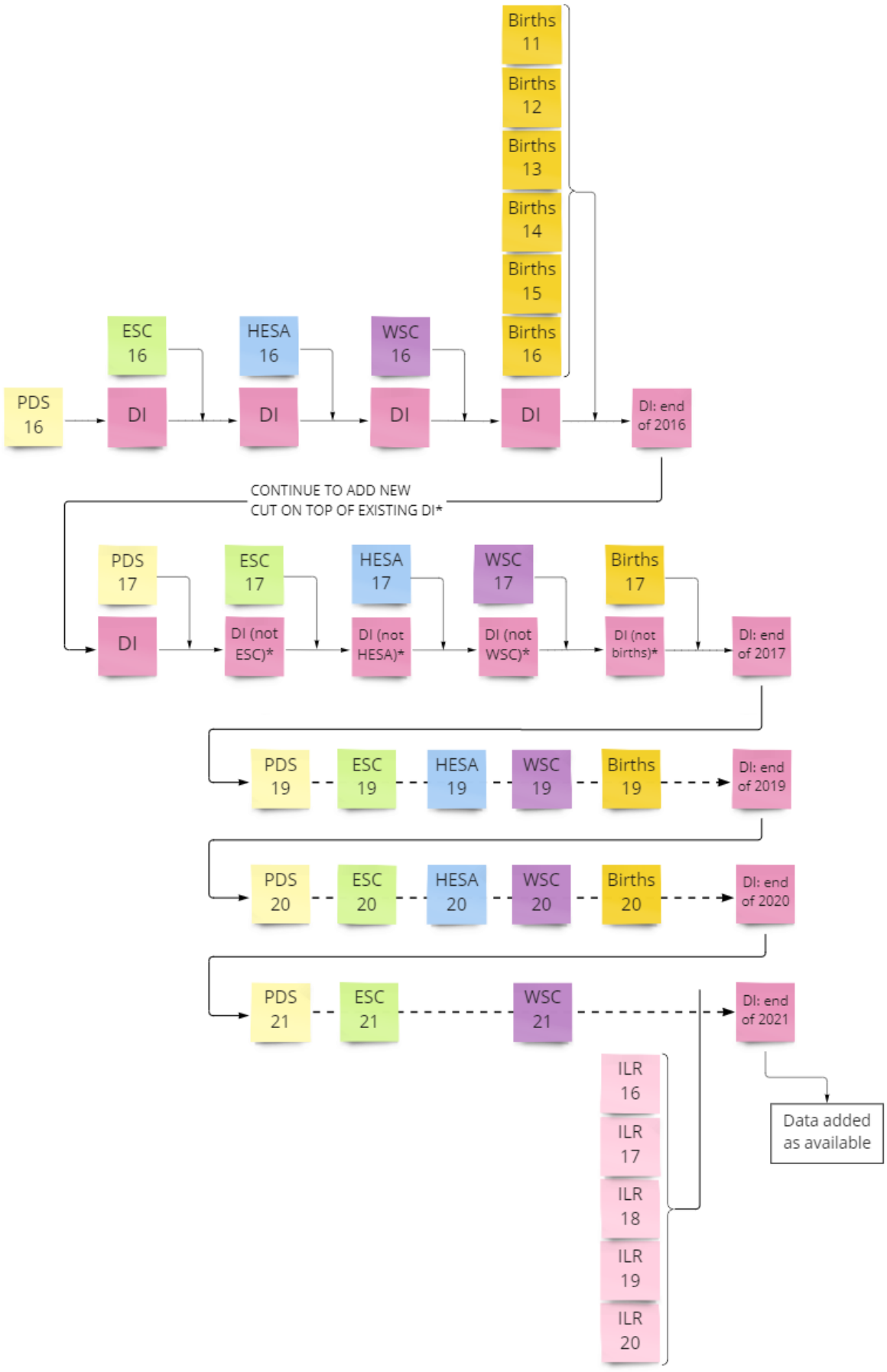

Figure 1: high level build for the DI v2.0, showing the order in which data are added.
HESA 2021 not available for v2.0; CIS data available 2012-2021.
*(*e.g. on addition of PDS 2018, any PDS 2016/2017 records in DI are excluded from the linkage stage; see Table 1 and footnote 16)*

The design – the process (Table 1): describes the stages that take place to build the DI, in order. It also highlights the key observations per stage that relate to statistical quality, and the recommendations that relate to these specific parts of the process to build the DI. Throughout this paper, the terms "cluster" and "clustering" relate to the cluster of records produced during the DI process, each of which receives an "ONS ID". The aim of the build is that each cluster will contain all records relating to an individual, across time and source.

Stage	Purpose	Key observations relating to quality	Recommendations arising
Standardising and cleaning	Data are made consistent in structure, such that all sets can proceed through the same pipeline	• Missing values are identified and coded as "None" • Both sex and gender fields are used to create a single sex/gender variable, for use as a common linking variable across sources[8]	• Ongoing identification and analysis of missing values particularly for linkage purposes • Test impact of amalgamating sex and gender variables for linkage[9] • Feedback loop to data suppliers to support continuous improvement
Adding UPRNS	UPRNs are assigned using AIMS[10], and are added to support delta and linkage stages	• Accuracy of how UPRNs are assigned in the DI is unknown • More generally, the quality of AIMS is important in ONS[11]	Investigate accuracy of UPRNs assigned Investigate characteristics of records that are not assigned UPRNs, to understand potential error in linkage as a result of this missingness
"Dedup/reject"	A small number[12] of records are removed from the DI ("rejected"), as they are expected to reduce linkage quality	Removing records could introduce bias	Examine rejected records, and test decision logic for rejection
"Delta"	An efficiency step, where records that duplicate linkage information[13] are removed from the data to be linked. This is so that effort is not duplicated during linkage.	Records removed at this step can be returned at the end of the DI process, and are not lost	Ensure that records are returned and available to users
"Combine linkage ID"	Records are clustered over time, within a source	Clustering over time is dominated by source ID (e.g., NHS number), and is supported by an exact match on PII	Investigate error in source ID, and impact on clustering[14]

[8] n.b. Gender (and not sex) is available on all sources except HESA, where a variable for sex (but not gender) is available. Therefore, both sex and gender fields are used to create a variable that can be used to link across all sources. This design choice was arrived at with the support of the Sexual Orientation and Gender Identity team, in ONS

[9] – n.b. the sex_gender_link variable is purely created and used for linkage; users do not receive this variable, and the fields for sex and gender in the data are never overwritten

[10] Address Index Matching Service

[11] The new AIMS methodology is going to be tested shortly and the results of this will be presented to MARP

[12] A few hundred – relatively few compared to the total number of records in the DI (~290,000,000)

[13] Linkage is based on source ID and PII, so records that have the same source ID and PII are duplicates in terms of linkage information, and linkage outcome

[14] e.g. same source ID assigned to two people, especially as this is more likely to happen when PII agree too (and therefore PII do not ameliorate source ID error)

Linkage	Linkage between a new cut of data and the existing the DI – this is where links between sources are made. Records are linked in consecutive exact, deterministic, and probabilistic matching steps. Links are generated between records, and used as evidence to link clusters. Probabilistic matching is achieved using Splink[15].	• Error/Bias in linkages are unknown • Probabilistic linkage scoring thresholds for accepting a link are hard-coded • Models for probabilistic linkage are partly generalised (e.g., one Splink model when linking any PDS data to the DI) • Linkage information is incomplete[16] • Records are partially excluded from deterministic and probabilistic steps[17] • Middle name is consistently available across sources, and so not much used in linkage rules (matchkeys), but is known to be useful for linkage with non-Western names	• Measure error and bias for every linkage[18] • Develop probabilistic linkage, including: improve threshold for accepting links, based on linkage results; investigate whether models are overly generalised; investigate whether Splink is properly implemented • Obtain complete linkage information, as it is harder to do QA without it, and as it would support future the DI (see below) • Test impact of excluding records from linkage • Develop methods such that extra effort is made to link single residual records to the DI • Improve linkage methods for non-Western names (middle name, and associative matching)
Reconciliation	Information connecting records over time and over source (e.g., from linkage) is used to cluster records.	• Clustering is additive – clusters can be joined but can't be split • As a result of clustering, ONS ID may be updated for some records in the DI (e.g., new data leads to a false negative being corrected) – this has implications for longitudinal analysis (see longitudinal research)	• Develop methods to evaluate clusters for false positives, and to split clusters • Reconcile changes in ONS ID with longitudinal analysis (see below)

[15] Splink is a PySpark package developed by Ministry of Justice, for probabilistic linkage at scale. It allows Fellegi-Sunter probabilistic linkage to be applied in a distributed system, and estimates parameters using the Expectation Maximisation algorithm. Further details can be found at "Splink: MoJ's open source library for probabilistic record linkage at scale" (found at https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/splink-mojs-open-source-library-for-probabilistic-record-linkage-at-scale)

[16] Linkage is record-to-record, but not all links are attempted or made or kept. Complete information would be generated if all record-record links were attempted and kept. Ways in which linkage information is made incomplete include: only residual clusters go through from exact to deterministic, and from deterministic to probabilistic stages; some records are excluded from linkage (see footnote 10); only the strongest record-record link between clusters is kept.

[17] Exclusion is of DI records that are from the same source as the new data. This design choice is to improve speed of linkage, and assumes that within-source links are adequately picked up during exact matching, or via links with other records ("inferred links", see below)

[18] per dataset added to the DI, not per step (exact/ deterministic/ probabilistic)

Recommendations: expanding on major themes

Metrics, testing, and research – measuring quality

Metrics: Across the design of the DI v2.0 we have identified and agreed a list of specific metrics that we believe are the first steps towards measuring quality for the DI. The full list of these metrics is available in the Annex (A1).

We recommend that these metrics be tested, in the first instance, on data made from linking the DI to Census 2021 and CCS 2021 (DI-CC). Those that prove to be good measures of the quality of the DI should be included in user reports for the DI.

We expect that metrics for linkage error, and "cluster metrics" will be particularly important in assessing DI quality. Examples of the latter may include:

- composition of clusters (number of records, and sources)
- number of source IDs in a cluster (e.g., >1 NHS number)
- correspondence of cluster and individual
    - more than one person in a cluster (false positive cluster)
    - more than one cluster for a person (false negative cluster)
- measures of variability within clusters (e.g., variation in age within a cluster)
- measures based on the structure of a cluster, drawn from graph theory, for example:
    - how connected is a cluster (e.g., many links between records within the cluster)?
    - how disconnected is a cluster?
    - the presence of any "bridges" in a cluster (where two internally well-connected clusters are connected by a single link)

We note that linkage to create the DI-CC will be restricted to one year of the DI records (2021, and 2020 for HESA), and to CCS areas. We expect and recommend that the DI should be linked to Census and CCS without a restriction on the year of the DI record. We expect that this would help us understand how to distinguish between error and change in the DI, how to use the DI for longitudinal analysis, and how to identify the presence and role of lag in the DI.

Testing: It is difficult to predict the impact of decisions made in the DI design on the final linktable, because the process for building the DI is complicated and involves many cuts of data. Therefore, we have recommended that the impact of various design decisions should be tested through simulation. For example:

- Does the order of adding data to the DI matter?
- What is the impact on the DI of excluding some of the DI records from deterministic and probabilistic linkage? (see footnote 16)
- Examine logic underlying the rejection of records
- Examine impact of missingness on rejecting records and linkage/ clustering
- Examine the role of older records in linkage/ clustering (i.e., perhaps details change over time and compromise linkage)
- Examine the trajectory of clusters over time – (i.e., see how they grow, when it is desirable, and when it produces false positive clusters)

The full list of these recommended tests is available in the Annex (A1).

We are developing with the DI team a “manual test” that can be used for the task of testing. It will draw on synthetic records, put them through the DI pipeline, and give back a synthetic DI.

Such testing will require synthetic data that adequately reflects the DI. This is tricky to develop because without understanding the DI quality better it is difficult to ensure that our synthetic records capture the quality issues that we want to test. Currently, our synthetic data is designed to reflect plausible “edge cases”, which were chosen for earlier work to stress-test the DI design (e.g., records for twins, to see under what circumstances the DI would successfully cluster them into distinct ONS IDs).

Therefore, an ongoing task will be to develop our data so that – when run through the pipeline – it produces a synthetic DI that looks sufficiently like the real DI. Metrics developed for quality purposes will prove useful here, by helping us to characterise the DI (e.g., number and type of clusters). We can then alter our pool of synthetic records so that the resulting synthetic DI shares those characteristics.

Testing will also require measures of quality to assess the impact of decisions. In the first instance we will explore the impact of a design decision on how well the DI clusters a given set of synthetic records. As quality metrics are developed, we can run them through our simulation to trial them prior to the DI-CC being ready.

The simulation work can support further recommendations. For example, we will use it to develop a synthetic DI with an altered pipeline that produces complete linkage information (see Table 1), which can then be used to develop a graph database prototype. It will also be useful for testing future proposed design changes proposed by the DI team, and so provide evidence towards design development.

Research: In particular, we highlight:

Longitudinal research – the DI is designed to cluster records for individuals, but does not create time series out of them. However, it does produce variables based on the date that ONS received records (“reference date”). This effectively creates a window of time over which details[19] for a given source ID are considered valid, but it is not equivalent to a time series as this “reference date” varies in definition across source.

Several non-trivial problems will need to be tackled before the DI can support longitudinal research:

- reference date varies in definition across source – e.g., School Census reference date indicates date-of-capture, but PDS reference date indicates when a cut was taken from the stock (see next)
- PDS records come from stock – therefore, a record may not necessarily indicate proof of presence, or contain details that are correct, but may simply indicate that a record has not been updated

[19] e.g., postcode, address, date of birth

- looking at records across time for an individual, we cannot differentiate between error and change[20]
- There is a potential security issue in assembling records for an individual overtime – e.g., it may become possible to identify when a person changes their gender

A particular challenge for longitudinal analysis is that the addition of new data to the DI will lead to corrections in clustering, meaning that ONS ID may change over time. Such alterations in how records are clustered would lead to a "ripple effect" on any longitudinal dataset, particularly in terms of household-level information. While the addition of new data may lead to errors being corrected (e.g., false positives or false negatives), it is also possible that error is introduced. Even if overall there is a reduction in error, changes in clustering will mean that outcomes from longitudinal analyses are unstable, meaning that reproducibility is difficult.

Further discussion of what longitudinal research should involve lies outside the scope of this paper, however, we recommend that this research be pursued.

Working with suppliers will become important, to ensure that the use of data remains acceptable, and to glean metadata that will support longitudinal analysis – for example, finding supporting fields that indicate activity on PDS.

Graphs research – graph databases are useful for storing information where the connections between items are as important as the items themselves. In particular, one may use a graph database to store both records and the linkage information that connects them.

By storing linkage information, graph databases allow flexibility to alter linkage without having to rerun the build of the DI; as the DI grows in size, this could lead to an important efficiency gain. We expect that this flexibility would make it easier to:

- integrate new sources
- optimise linkage/ build
- test out alternative designs at scale (e.g., alternative linkage error thresholds), without having to re-run linkage
- resolve clusters (e.g., splitting, reforming clusters on addition of new data)

These benefits are particularly important for data sets that are made out of multiple linkages, like the DI, and RDMF ("composite" data). Therefore, we recommend that research into using graphs to support the DI development also be pursued.

Besides these areas of research, the National Data Strategy[21] calls for further research into Machine Learning to improve linkage, and the trial of Bloom Filters for hashed linkage.

[20] Arguably this work belongs to projects such as the production of SPDs, which addresses the problem using inclusion and location rules.

[21] https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/joined-up-data-in-government-the-future-of-data-linkage-methods

Usage and Governance – Using the DI

Usage: In order for the DI to be used for making statistics, users need to measure the quality of their outputs and use the DI such that they achieve the best quality results. We believe that to do this, they will need to understand what the DI is, what is its quality, and how that quality affects the quality of their outputs.

To make this possible, we recommend that a support service be developed, and that it should include:

- documentation about the DI:
    - version-specific information (e.g., changes since the last version, data in the build, Splink version and models)
    - quality metrics
    - data dictionary/ glossary
    - examples of how to use the DI and to show how it is built (e.g., how synthetic records are clustered)
- the development of "use cases", where the teams behind the DI and MDQA work with users to help them use the DI for their own analyses

Further to this, we recommend that flags for deaths and migration be developed, and that the method behind them be made available to users, too.

In the Annex we provide "Guidance on Usage", which offers initial guidance for the user based on what we've learned from this first iteration of MDQA.

Governance: We recommend that technical governance be set up for the DI, led by Data Architecture and supported by MQD. This will provide a necessary link between the DI development, the ongoing work to measure quality, and the task of communicating both to users.

We recommend that the role of governance should include:

- the presentation and discussion of design changes, which could be presented alongside an assessment of their impact on the DI
- discussion of how the DI development can meet business needs
- overseeing the implementation of any improvements identified following analysis of the DI-CC
- overseeing a beta testing phase (including establishment of success criteria) prior to roll-out so analysts can test and report back on use of the DI
- discussion of statistical and quality impacts of design decisions

Challenges

Hashed data – It is unclear when unhashed data will become available to replace hashed CIS data in the DI. Therefore, at present, we assume that measuring the DI quality will mean measuring the quality impact of hashed linkage on its build.

After the main build has been completed, CIS data are linked at record-level to each other source in the DI (CIS to PDS, CIS to ESC, etc). The method for hashed linkage – Derive and Conquer – has been quality assessed for CIS-PDS 2019. To quality assure hashed linkage in the DI, we would need to extend this assessment to the other

sources, and for all years. This will require collaboration with DWP, and agreement from DWP that we may receive multiple samples of CIS records, unhashed[22].

Moving from record-level links to the cluster-level, hashed linkage has a second impact on the DI through the production of "inferred links". These occur when two records in the DI are clustered not by linking directly to one another, but by both linking to a third record. In theory, we could assess the impact of inferred links on clustering in the DI through the DI-CC; however, it is not yet clear if we may inspect inferred links in conjunction with Census data under our existing data-sharing agreement with DWP.

"One size fits all" – We do not yet know if the DI v2.0 will support all types of analysis equally well – e.g. both point in time and longitudinal. A clearer understanding of quality (i.e., through metrics), and use case work will provide evidence here, as will research into longitudinal analysis. In particular, it will be important to track quality from the DI through to outputs, and to examine quality across a range of outputs.

Automation – This is essential, to allow timely production of the DI and assessment of quality. Also, a key use of the DI in the future will be the Demographic Index Matching Service (DIMS), which will link users' data to the DI. Therefore, the demand for quick and accurate linkage at scale is only going to increase.

We expect that clerical review will be required to assess quality in the DI – but this work is notoriously expensive in terms of time and resource. Therefore, a balance must be found between speed of production and production of a high-quality the DI whose quality is known.

Here, machine learning approaches may be helpful, for example active learning could aid the selection of records and clusters for review. However, we also note that automated methods are optimised for maximum coverage, which can lead to marginalised sub-populations being poorly served; therefore, any such methods will need to be carefully examined for bias.

Ongoing validation: The production of the DI-CC will be vital for assessing the DI v2.0 quality. However, we expect that ongoing validation exercises will be important, perhaps through linking surveys to the DI.

Flexibility, and optimisation: We expect that graph databases would support flexibility in the DI build, leading to faster integration of new data, and greater ease of altering and testing design at scale. See research and graph databases, above.

Resource requirements: Many of the challenges above will require time and resource in order for us to tackle them. Currently, MQD resource will allow us to support the DI-CC linkage, research into metrics and graphs, some testing, and support of governance. All other activities are currently backlogged. It will be an important job for the new governance structure to ensure that available resource is balanced against the highest priorities.

[22] It may be possible to reduce the total amount of unencrypted data required by reusing samples.

Future trajectory and Conclusion

The recommendations set out in this paper constitute the “next steps” that we think should be taken to realise the overall goal of MDQA and the DI.

At the outset we described this goal – to support development of the DI such that it can be used to produce high quality statistics, with known statistical quality, across multiple analyses. The current work represents a first step towards it and has helped us to identify further steps on the path. We present a summary of these in Table 2.

Conclusion: Through this first iteration of MDQA for the DI, we believe we have identified specific and actionable steps that can lead to measuring quality in the DI, and to realising its potential in ONS.

To measure quality, the highest priority work is to develop and test metrics for the DI quality and to test the DI design through simulation. And in order for the DI to be used – and used well – work to support users is vital.

Lastly, putting in place the right governance will be important for giving the DI sufficient structure as it matures. This will help to achieve the right balance between business needs and development, and will help to foster a better understanding of the DI – and of the opportunities offered by it – in ONS, allowing users, those developing the DI, and those engaged in the ongoing task of articulating quality, to move forward together.

Table 1: future steps for ongoing MDQA of the DI

Stage	Details	MoSCow rating	Estimated size
Assure the DI design (v2.0)	Clarify the DI v2.0 design, including: build, mechanisms, design decisions, assumptions Use design to identify metrics and further analysis for developing our understanding of quality	Must	L[23]
Establish a governance structure for MDQA and the DI	Support ongoing maturation of MDQA, ensure the DI and MDQA stay in step, and review requirements with stakeholders	Must	M
Metrics and further analysis	Metrics and further analysis, with a view to creating quantitative measures of quality in the DI – as proposed throughout this paper	Must	L
Evidence/ Data to validate the DI	High quality linkage to high quality data is required to validate quality in the DI (e.g., the DI to Census/ CCS Linkage)	Must	L
Review and explore existing quality frameworks, in order to develop the DI quality framework	To include Longitudinal Error Framework[24], Stats NZ framework for admin data[25]	Should	M
Clerical review of all linkages in the DI, including hashed linkages	Proposed in recommendations to review exact/deterministic matchkeys (MKs), and to review the use of Splink (above).[26]	Must	L
Support novel research to develop the DI	Longitudinal analysis, graph databases, machine learning	Must	L
Assess quality for a selection of use cases	Work with users to produce specific outputs/statistics from the DI, then evaluate the quality of outputs	Must	L
Raise the question of a wider inference framework	Development of methods to support production of statistics from the DI (composite data) is outside scope of MDQA, but is very important and related to the task of measuring quality	Should	M
Raise the question of liaising with data suppliers	To improve quality in source data To gather metadata that can inform data use in the DI	Should	M

[23] This first iteration of MDQA completes our assessment of the design of the DI (v2.0). Ongoing work will be required to test various aspects of the design (see Annex A1), and to ensure that design changes are logged and examined.

24 Longitudinal Error Framework, found at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsworkingpaperseriesno19anerrorframeworkforlongitudinaladministrativesourcesitsuseforunderstandingthestatisticalpropertiesofdataforinternationalmigration

[25] Stats NZ framework for admin data, found at: https://www.stats.govt.nz/methods/guide-to-reporting-on-administrative-data-quality

[26] Although ONS may move away from using hashed data in the future, current the DI builds include hashed linkage. Therefore, the need to quality assess hashed linkage remains germane.

Annex A1: List of Metrics and Tests (taken from main paper, Appendix A4[27])

"Part of Design" refers to stage of design, and/ or to the part of the main paper where the recommendation arose

Metrics: single numbers – may relate to the process of the DI (has the code run successfully), or to statistical quality

Cluster metrics: require development – metrics that relate to the composition or structure of clusters, as described above; these also require data validation to successfully identify which metrics indicate poor or good quality clusters

Tests[28]: simulation of the DI pipeline/build, especially to send through specific example records (scenarios) to test the effect of build on outcomes

Clerical review: inspecting samples of records

Further analysis: more digging, beyond a single metric or use of data examples, to answer a broader question

Data validation: high quality linkage to a high-quality data source, for the purposes of validating the DI quality (e.g., the DI linked to Census/CCS (DI-CC))

Research: expanding further analysis, or developing a novel approach

Part of design	Description	Type of work suggested
High level build	coverage vs. error – what is the impact of adding data?	Tests
	Test impact of using sex and gender to create a single field for linkage, on linkage quality	Tests
	Does order of build matter?	Tests
	What is the impact on final DI clusters of restricting linkage (excluding existing DI records from deterministic and Splink steps, if they are from the same source as the new cut)	Tests
	Hashed data: Exploration of inferred links, to evaluate role of CIS records in connecting other records	Data validation
	Hashed data: QA of CIS links (generalisability of Derive & Conquer hashed linkage method). *Requires samples of data in the clear*	Clerical review
Staging, cleaning, standardising	Proportion/ number of records in the DI that receive a UPRN (*any* threshold of AIMS confidence)	Metrics
	Proportion/ number of records in the DI that receive a UPRN (that meet threshold of AIMS confidence)	Metrics
	Proportion/ number of records in the DI that do not receive a UPRN, but could (i.e., PDS, ESC)	Metrics
	Test impact of error in UPRNs on final clusters in the DI	Cluster metrics, Tests

[27] Taken from our previous paper, available on request

[28] Sometimes described as "scenarios" or "data examples"

	Measure/ explore error in UPRNs How many of the UPRNs assigned to DI records are correct?	Clerical review
Dedup/reject	Count of rejected records	Metrics
	Distribution of characteristics for rejected and almost-rejected records (incl. details about data – e.g., which source)	Metrics
	Examine logic for rejecting records and its impact on final DI clusters	Tests cluster metrics
	Examine impact of missingness on rejecting records, and impact on final DI clusters	Tests cluster metrics
	Examine sample of rejected records: were they correctly rejected, can they be re-introduced?	Clerical review
	Examine sample of records that were *nearly* rejected: should they have been?	Clerical review
	Find bad clusters in final DI and identify whether any of these could be fixed with improved dedup/reject logic	Data validation
Delta	To indicate successful unpacking of replicate stack – counts of records, and distributions of characteristics	Metrics (process)
	Count and characteristics distributions of duplicates, per year and source	Metrics
	Count and characteristics distributions of identicals, per year and source	Metrics
	Test impact of identical/duplicate definitions (do they need to align)	Tests, cluster metrics
	Test the assumption that name, dob, sex, and postcode are the only variables that underpin linkage (definition of identical)	Tests
	continue work to produce change variables, and confirm that they are operating as intended, and robust to errors/missingness	Tests
	If CR1 is to be used for more than linkage efficiency – explore options for improving comparison/ creation of CR1	further research
	Identify when an upsert[29] is an error, and when it is an update, with a view to improving quality in the DI	data validation
Combine linkage ID	Inspect linkage ID clusters, both edge cases and as identified through data validation	clerical review, cluster metrics, data validation
Linkage	Count and percentage of records linked after linkage of a cut of data	Metrics
	Count and percentage of records left unlinked at the end	Metrics
	Number of duplicate links made by MKs	Metrics
	Total number of links made by MKs	Metrics
	Number of unique links discarded in the process MKs	Metrics

[29] DI jargon: an upsert is a record in new data that is being added to DI, where the source ID is not new. That is, it is assumed to be someone who is already captured in the DI but who has a change in PII – i.e. refreshed linkage information.

	Number/proportion of records that go through to deterministic and Splink stages	Metrics
	Characteristics of records and type of clusters that go through to deterministic and Splink stages	Metrics
	How many many-to-one links are made by Splink?	Metrics
	Review of MKs identify duplicate MKs, examine generalisability of method, examine use of middle name for MKs, the use of key order to resolve conflicts, stress-test with difficult linkage examples – impact on final DI clusters	Tests, cluster metrics
	explore order or balance of different linkage steps – impact on final DI clusters	Tests, cluster metrics
	generate full linkage information, and use it to explore alternative clusters – particularly in the cases where final clusters look incorrect[30]	further analysis, cluster metrics
	Review MKs: inspect unique links that are discarded, and weaker links that are kept – should they have been kept (would they have made bad clusters better in final the DI?)	clerical review, cluster metrics
	Clerical review to sample scores and confirm thresholds for accepting links in Splink step	clerical review
	Consider/develop a more flexible approach to automatically accepting scores in Splink (beyond a fixed threshold)	Further analysis
	Test generalised models for Splink: examine linkage quality, particularly early on in the DI build	Tests, further analysis
	Investigate many-to-one links made by Splink and review logical basis for accepting such links	clerical review
Clustering	Count and percentage of records/clusters that have had their ONS ID changed as a result of the reconciliation process[31]	Metrics
	Develop metrics to identify when clusters are not 1:1 with individuals, (based on identifying “bad” clusters through edge cases and data validation)	Cluster metrics, data validation
	Develop methods for splitting clusters where more than one person is contained	Further analysis
	Research into graph analysis using information about clusters (explore cluster metrics that are based on cluster structure)	Further analysis
Residual records	Count and percentage of residual records after each cut is added	Metrics
	Further analysis of key characteristics for residual records: single year of age, sex, geography, data source, reference year, nationality, country of birth, and ethnicity	Metrics
Missingness	Per data cut: prevalence of item missingness for key linkage variables, including patterns of missingness	Metrics
	Per data cut: characteristics of individuals with missingness	Metrics

[30] Also relates to graph databases, as we expect this type of analysis to be much faster using a graph to store linkage information

[31] n.b. these metrics include “count and percentage of clusters merged during processing”

	For final DI linktable, number and characteristics of clusters with missingness	Metrics
	Introduce missingness to synthetic data to explore the impact of missingness on final DI clusters	Tests cluster metrics
	Investigate impact of missingness on individual blocks	Tests cluster metrics
	Further development of how missingness is identified during standardisation	Further analysis
Duplication	Number of false positive clusters	Cluster metrics
	Number of false negative clusters	Cluster metrics
	Evaluate clusters 1:1 with individuals by linking to high quality data and examining edge cases	data validation, cluster metrics, clerical review
Data over time	Investigate the trajectory of clusters over time (incl. 1:1 correspondence); in particular, with respect to the addition of data – to join up false negative clusters and to create false positive clusters	Tests, cluster metrics
	Explore the role of older records in the DI in the build process and on the final DI clusters – e.g., links involving older records could be weighted down in comparison to links involving newer records	Tests, cluster metrics
	Investigate production of time series from the DI	Research
One size fits all	Research into graphs, and graph analysis: ways to store and use linkage information in the DI: allow “tuning” of the DI for different use cases (n.b. DIMS, too), and support addition of data to the DI	Research, cluster metrics
Characteristics	Include amongst cluster metrics those that investigate variation in name (e.g., family members clustered by mistake)	cluster metrics
	Describe and compare distributions of characteristics for final build of the DI (and throughout)	Metrics, cluster metrics
	Investigate the Bias Analysis Tool, developed by the Linkage Hub – try application to the DI, and consider developing it to handle composite data	Further analysis
Deaths and migration	Count and percentage of records where people are thought to have died, by single year of age, sex, geography, and source	Metrics
	Number of death records that did not link to the DI (e.g., no NHS number)	Metrics
	Number of records where people are believed to have emigrated, by single year of age, sex, geography, and source	Metrics
	Investigate clusters where a death has been flagged (linking on NHS number)	clerical review

Future the DI	Investigating the role/use of surveys to support the DI – e.g., to validate/check the DI on a rolling basis	Research
	Expanding the DI to include and use relational data between people – e.g., fields for next of kin	Research
	explore non-greedy optimisation for the DI (ML) – may fit with graphs work	Research
	ML for supporting efficient/strategic ongoing clerical review	Research
Top priorities	Further research to build on the existing DI to Census/CCS linkage plans for PMST, to support full data validation of the DI across time and geographies	Data validation

Annex A2: Guidance on Usage – example guidance

What is the DI? the DI v2.0 is a composite data source, a patchwork of many cuts of admin data, linked across source and time. The DI and MDQA are relatively young, and still developing; they will reach maturity with usage, the development of quality measures, and supporting governance.

Therefore, we urge users to engage with the DI, and grow their understanding of it, in order to use it effectively. In particular, this means consideration of quality – not just whether analysis yields numbers, but whether those numbers have a high statistical quality.

The following are key characteristics that analysts should know about, when using the DI v2.0:

1. Records are clustered

The main aim of the DI build is to cluster records and assign each an "ONS ID". Each cluster stands for a person.

2. The DI is not a simple population dataset

Since the DI is composed of clusters, this means that it is not a simple record-per-person dataset.

3. The DI is not ready "out-of-the-box"

Within a cluster there may be variation among values – e.g., a variety of ages – for this reason, users are expected to apply their own business rules to the DI, in order to resolve clusters such that they can be used to make statistics.

4. Analysts should keep in mind how records are captured

Across the sources in the DI there is variation in how records are captured – i.e., generated by admin providers, and passed to ONS. This affects how we can interpret what records stand for and should be considered by analysts.

In particular, whilst records from ESC, WSC, and HESA always indicate evidence of an individual at time of capture, cuts of PDS data are snapshots taken from a PDS "stock". A summary of capture mechanisms, and the implication for records in the DI is provided in the Table below.

Source	Capture mechanism	Implication for the DI
School Census (ESC, WSC)	Every January, students are recorded on School Census	A school census record indicates positive proof of presence and details, at the time of capture
HESA	At the start of every academic year, students are captured on HESA	A HESA record indicates positive proof of presence and details, at the time of capture
PDS	Every cut of PDS is a snapshot of the "stock" of PDS records. A new record is generated when someone registers with a GP. We do not	Records persist in the absence of update or cleaning – i.e., activity on behalf of the individual or the provider. For example, if someone registers in

	believe that PDS stock is regularly cleaned.	2016, that record enters the "stock", and given no further activity, it will be resupplied in the cuts for 2017 and 2018.
Births Register	Births are registered at or near to date of birth	A birth record indicates positive proof of presence and details, at the time of capture
ILR	As for HESA	An ILR record indicates positive proof of presence and details, at the time of capture

5. Work to establish what users receive as the "DI" are ongoing

Work to establish quality (MDQA) has focussed on the build of the DI v2.0 "linktable". This is the full dataset that results from the process for building the DI, and still contains the Personal Identifiable Information (PII) used for linkage. For disclosure reasons, users will receive some version of the linktable that has been redacted to reduce sensitivity – for example, with PII removed. Discussions to establish what users will receive are ongoing.

6. The DI will support many analyses, but not all

DI v2.0 contains the following source identifiers: NHS number, pupil matching reference (School Census, England and Wales), HESA student number, and ILR student number. Therefore, if an analysis requires linkage across any of these identifiers, it will be supported by the DI v2.0.

Due to DWP sharing agreements, National Insurance number (NiNo) is currently only available to ONS encrypted. However, it is still possible to link data that has NiNo to the index, if that data has its NiNo similarly encrypted.[32]

7. The DI is not yet ready to support longitudinal analysis

This is because reference dates refer to when cuts of data are received, and do not necessarily indicate positive proof of an individual's presence on that date; it is also because change and error cannot be disambiguated. A key recommendation from MDQA is that research into how to use the DI for longitudinal analysis be pursued.

The DI team have developed fields to indicate when records, and details such as address, are current (valid).

8. There are no variables to tell the difference between an error and a change

Currently, we cannot tell when a record contains an error. Similarly, when examining records belonging to an individual (as defined by a cluster), we cannot tell where variation indicates real change and where it indicates error. Analysis of the DI-CC[33] will support research to understand and identify error and change in the DI.

[32] The DI team will be able to carry out this encryption and linkage on behalf of users, as part of the Demographic Index Matching Service (DIMS)

[33] the DI linked to Census and CCS data

9. Work to establish quality is ongoing

The DI team and MQD are working together to assess the quality of the DI, and to support users. This work includes:

- quantitative quality metrics to measure quality in the DI
- development of death and migration flags for records
- creating user documentation to report on quality and provide essential information for using the DI
- development of use cases, where the DI and MQD teams collaborate with users to join the dots between the quality of the DI and the statistical quality of their outputs
- establishing a governance structure to report on how the DI and MDQA develop and progress

10. A better understanding of quality will be supported by linking the DI to Census and CCS

By linking the DI to Census and CCS (DI-CC), we will be able to validate which of our quality metrics are good measures of quality in the DI. Similarly, we will be better able to answer questions about change and error. Linkage work is underway, and analysis of the DI-CC is due to take place in Autumn 2022.