Proof of Concept – Address Centric Admin Combined Intelligence Dataset (ACID)

Contents

Introduction 2
Methods 3
Data access 3
Table 1, Potential datasets for use in ACID. 3
Base data 3
Table 2, Datasets available and those used in the creation of ACID. 4
Summary of Modelling Approach 5
Data Linkage 5
Modelling Covariates 6
Modelling the data 7
Table 3, Modelling summaries 7
Evaluation 8
Future Steps 8
References 9
Appendices 10
Appendix 1: ABPE vs. ACID 10
Appendix 2: Match Keys 11
Best matches by match key 12
Appendix 3: Modelling Variables 13
Appendix 4: Logistic Regression Detail 14
Appendix 5: Random Forest Model Detail 15
Appendix 6: Assumptions and Limitations 16

Introduction

We know from previous census experience both here and abroad, and from the coronavirus pandemic, that we need to be prepared for the unexpected. The Address Centric Admin Combined Intelligence Dataset (ACID) is a proof of concept project that has been set up to enable work on census low response contingencies based on a number of possible scenarios (ONS, 2020b).

These scenarios include:

1. expected individual questionnaires not returned (as found in the 2018 New Zealand census)
2. localised census count issues (as found in the 2016 Canadian Census with Fort McMurray)
3. broader census count issues (such as missing the overall census quality targets)
4. population subgroup issues (such as higher than expected non-response for a particular ethnic group or community).

ACID could be used to help address the first three potential scenarios but due to likely biases in the admin data would not be suitable for addressing the fourth scenario.

This project aims to investigate how to compile intelligence from various admin data sources linked at address and individual levels to explore the relationship between admin and census records at address level and help make an adjustment for non-responding households. It also aims to identify non-responding addresses on the census address frame, and to identify addresses where there is good evidence of current occupancy and to understand their likely basic household structure. Based on the exploratory proof of concept we will determine whether ACID is a viable product to help respond to the contingency scenarios.

There are various challenges such as the availability and timeliness of admin data, and over-coverage and under-coverage in the admin data itself. We are aware that our method is unlikely to remove all instances of over-coverage at address level; that traditional Census Coverage Survey (CCS) (ONS, 2016) covariates are not available in admin data; and we need to consider how ACID can be used alongside the traditional census processing. The proof of concept makes use of the available data which includes an address identifier (Unique Property Reference Number [UPRN]).

ACID would be run alongside the census but is to be used in a distinct manner. It would be the last resort if the standard design were at risk of fundamental failure and the standard coverage estimation and adjustment approach were in need of support. It would help to estimate the population in localised areas and be used to support household non-response adjustment in both census contingency and standard design. In addition to these specific uses it would also act to further inform longer term design for future admin-based population estimates (ABPE) and household statistics (details on how ACID differs from the ABPEs can be found in Appendix 1).

However, the intention was not to use the proof of concept ACID to create an estimation in the rehearsal but instead to test and demonstrate the data flows and infrastructure necessary for future use. We developed the following targets:

1. Minimal viable product: project area set up with Data Engineering linked data in it
2. Medium viable product: as above, but with data linked by address but no modelling
3. Full viable product – delivery of functioning product including constructed data frame and some completed modelling.

We currently feel that we have delivered somewhere between the medium and full viable products and are, at this stage, seeking feedback on the project so far and suggestions for possible modelling approaches. The methodology and approach taken for modelling will continue to be reviewed once the proof of concept version of ACID is delivered; with iterative improvements to be made to the methods thereafter.

Feedback from ACID fed into the Processing and Outputs Census rehearsal evaluation report. Initial progress on the ACID proof of concept has been reviewed by experts across ONS and is being presented to the Methodological Assurance Review Panel (MARP).

Methods

Data access

Initial meetings were held with a range of stakeholders to determine which admin data sources would be of most use for the development of ACID. The resulting list was used as a guideline for the actual data that was requested after consideration of availability, timeliness, and appropriate geo-referencing of the datasets.

These provisional datasets are included in table 1 and are colour coded according to whether they were requested and are currently in the project workspace. Note – throughout the development of this project certain datasets have been requested and then dropped according to whether they met our needs.

Table 1, Potential datasets for use in ACID.

Available and in project	**Unavailable and/or not in project**
Personal Demographic Service (PDS)	Births
AddressBase (Address Frame)	Census 2011
English School Census (ESC)	Labour Force Survey (LFS)
Welsh School Census (WSC)	DVLA
Higher Education Statistics Authority (HESA)	Annual Population Survey (APS)
Electoral Register (ER)	Royal Mail
Customer Information System (CIS)	Exit Checks
Benefits and Income Dataset (BIDS)	Self-Assessment
Deaths	Pay as You Earn Real Time Information
Hospital Episode Statistics (HES)	
Migrant Worker Scan (MWS)	
Council Tax (CT)	
Valuation Office Agency (VOA)	
Demographic Index (DI)	
2019 Address Frame	
2019 Census Rehearsal Responses	
2019 Response Management (RM)	

Base data

The base for this project is the 2019 Census rehearsal address frame. This frame consists of 331,359 distinct UPRN spread across 4 local authorities covered by the 2019 Census rehearsal. The Local Authorities used were Carlisle, Ceredigion, Hackney, and Tower

Hamlets. The proposed method for using the rehearsal address frame was to assign as many people as possible to a property, through a Unique Property Reference Number (UPRN), from the admin data, before using more activity-based sources to indicate 'signs of life' at every property.

The primary source of information for this project was intended to be the Demographic Index (DI). The DI is an internal statistical dataset that seeks to uniquely identify every person who has registered or interacted with the following systems: PDS, CIS, HESA, ESC and WSC. The initial proposal was to link the DI to the 2019 Census rehearsal address frame via unique property reference number (UPRN) with the aim of giving us information about the characteristics of individuals at a household (UPRN) level.

However, the way the DI has been constructed means that there is no specific information about the individuals in it and instead, it just contains linked unique identifiers from the individual sources. As a result, for this proof of concept work we chose to use PDS records to populate the UPRNs in the Rehearsal address frame. We chose the PDS because everyone resident in England and Wales can register with a GP and should therefore have good coverage with the exception of a small number of residents who use private health care exclusively.

Having taken the PDS as our primary data source, we have linked this data to the responses from the Census Rehearsal conducted around October 2019. We have also included a number of other sources of administrative data such as the English School Census, the Electoral Register and Council Tax data. The intention with these is to produce a logistic regression model that seeks to identify the characteristics of an admin data profile associated with the PDS having the correct individuals in the correct addresses.

Table 2, Datasets available and those used in the creation of ACID.

Available and in project	**Unavailable and/or not in project**
Personal Demographic Service (PDS)	Valuation Office Agency (VOA)
AddressBase (Address Frame)	Hospital Episode Statistics (HES)
English School Census (ESC)	Welsh School Census (WSC)
Electoral Register	Higher Education Statistics Authority (HESA)
Council Tax (CT)	Migrant Worker Scan (MWS)
Deaths	Customer Information System (CIS)
Demographic Index (DI)	Benefits and Income Dataset (BIDS)
2019 Address Frame	2019 Response Management (RM)
2019 Census Rehearsal Responses	

Table 2 contains a summary of the datasets we have used in preparing for this modelling approach. The majority of those we have not used lack the necessary address information (UPRN) to include them within our final dataset. They would, if this information were present, add value but we cannot use them for the time being. The VOA and Response Management data do contain UPRN but are unable to tell us anything specifically about the individuals in each property and we have excluded them from our dataset for this reason.

Summary of Modelling Approach

Our aim with the modelling is to link, at a record level, the PDS and Census Rehearsal Responses, to identify the administrative data records who also appear in the Census. We would then derive a number of covariates from the PDS and other administrative data sources to create an individual's administrative data 'profile'. These would then be used to construct and test a logistic regression model where having a confirmed link between PDS and Rehearsal is the outcome variable. The results from this model would then be applied to individuals from the Rehearsal Local Authorities which did not respond to determine who, based on their profile, we think is in the correct location according to the administrative data.

We are looking to use a model-based approach as it will allow the data to form its own conclusions as to what is and isn't associated with being in the correct location. Whilst it would not be difficult to make assumptions about which covariates would be most associated, by using a modelling approach we let the data drive these decisions and reduces the risk of error or bias being introduced by our assumptions.

Summary

- Link Census Rehearsal responses to the PDS
- Create covariates to feed into regression modelling
- Train model for each Local Authority
    - Possibly split data further – adult/child for example
- Test model on remaining subset of data where we have Rehearsal responses
- Apply results from model to non-responding households

Data Linkage

For the data linkage, we only used PDS data for UPRNs which had submitted a Rehearsal response, this ensured that the process gave us a better understanding of the coverage and accuracy of the administrative data in relation to Census responses.

To conduct the linkage between the PDS and the Census Rehearsal, we used a series of deterministic match keys, 21 in total. More details about the match keys and the numbers of best matches for each one can be found in Appendix 2. All the match keys require there to be an exact match on UPRN between the two data sources. This was important for this research as we needed to ensure that we were understanding the coverage patterns within each property. For all other linkage variables, such as name, sex and date of birth, allowances were made in the match keys for some amount of error.

At present, we are conducting analysis to understand the structure of the populations in the data. We want to be sure that we understand who the PDS thinks lives in each LA, who responded to the Census Rehearsal in each LA and also the characteristics of the matched and unmatched populations. This is key to understanding any biases that the data may present us. We also clerically reviewed a portion of the matches, particularly focussing on the more 'relaxed' match keys, to ensure that we were happy with how the linkage process had gone.

For those records where we did find a match between the PDS and the Census Rehearsal, we flagged these in the base PDS data and this then carries through the rest of our work,

including into our final dataset which contains a number of covariates we have derived for modelling.

Modelling Covariates

The covariates for our model have come from a variety of data sources. This has proven to be particularly important as we do not have all data for all Local Authorities. For example, for the children in England, we can use the English School Census, but the equivalent data for Wales does not contain UPRN. We also do not have access to Council tax data for all 4 Local Authorities used in the Rehearsal exercise. Further detail on the outcome variable and covariates can be found in Appendix 3.

From the PDS, we have used the Reason for Removal information which indicates if someone should be removed from the data and, if they should, the reason either the individual or the GP feels they should be removed. Initially, this information is all stored within one non-ordinal, categorical variable so we have transformed the data to have it as a series of dummy variables. We would expect that some of these would be more associated with the likelihood of finding a match.

The PDS also contains a number of date variables. These relate to times when information in the central system has been updated and cover things like address changes or moves between practices. We have calculated the time difference between these and the Census Rehearsal date – 17th October 2019 – and our hope in this instance would be that the less time has elapsed between the two, the more likely the data is to be correct. Obviously, there could be cases where information has not been updated for many years and is still correct, but this is why we are feeding all of the information into a model rather than manually selecting what information is and is not correct. Additional research is planned to investigate the effect of time between data cut and reference date.

As well as the PDS, we have also included information from Council Tax records and the Electoral Register for our model. As we do not have enough variables in these two data sources to perform record level linkage and look for address conflicts, we instead decided to link the data using the UPRN. This meant that we were able to compare every surname the PDS contained in a property to every surname the, for example, Electoral Register contained for that same property. To compare these names, we used a method that included calculation of the Levenshtein edit distance to give us a score between 0 and 1 for how similar names were. A perfect match would give us a score of 1, decreasing as the names became more and more dissimilar. This is similar to the method used in our deterministic match keys as well as in the construction of the Demographic Index.

Finally, for children in the three English LA's, we repeated the name similarity comparisons using the English School Census. This does not include children who attend independent schools but this is a relatively small percentage (6.5% of all school children in the UK [ISC, 2020]). This should provide us with much more information about the children in properties as they will not be represented on either the Electoral Register or Council Tax data. We have also flagged whether or not the PDS and ESC have children of the same age in each property, again providing us more information about how accurate the administrative data is. This will be an interesting test as we will be able to compare the English and Welsh LA's to see how beneficial the School Census data is to our model. The WSC data that ONS have access to do not include UPRN and were therefore not able to be linked.

Modelling the data

The overall aim of any model we use is to answer the question, “conditional on getting a Census response, does the Census data person record match an administrative data person record.” Our intention for this work is to assess a variety of models, using the match/no match binary information as its dependent variable and the various covariates as the independent variables.

Conceptually, the model would look at the admin data profiles of those in properties which responded to the Census Rehearsal and calculate coefficients that determine the likelihood of a PDS record also being in the Rehearsal responses. We are consulting with colleagues from Methodology as to exactly how we should build the model and can use the results of these to meet our aims and look forward to making further developments.

In the meantime, however, we have already had a first attempt at training and testing two different models on the data for Ceredigion. We have considered a Logistic Regression and Random Forest model so far. For both of the these, we have assessed the model with all of the covariates and also reduced versions, taking out the insignificant covariates for the Logistic Regression and the Reason for Removal information for the Random Forest Model. All four have, so far, provided positive results as can be seen below in Table 3. The Logistic Regression models have been better at correctly identifying negative cases while the Random Forests are better at identifying positive cases.

Table 3, Modelling summaries

		Specificity	Sensitivity
Logistic Regression	**Full Model**	0.95	0.83
	Reduced Model	0.93	0.83
Random Forest	**Full Model**	0.83	0.88
	Reduced Model	0.81	0.88

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Further detail into the outcomes of the models can be found in Appendices 4 and 5 but in every case, the surname comparisons are the most important covariates. In the Logistic Regressions, a perfect name match increases the likelihood of being a match by over four times for the Electoral Register and over three and a half times for Council Tax. In the Random Forest models, these covariates are also the most important features.

We are continuing work on understanding these results, as well as seeing how they differ when similar approaches are applied to the data for different Local Authorities. There has, however, been very little variance in these figures across multiple runs of the same methodology. This suggests one of two things: that the data is not susceptible to sampling biases or that it is always susceptible to the same ones. We will be conducting analysis into the characteristics of the matched and residual populations to see how different they are, as well as comparing them to the overall Local Authority administrative data populations to see if we can identify any likely biases in the data.

The modelling process also has some areas where we have needed to, and will continue to need to, address issues within the data. We originally had the date differences in number of days however this made it incredibly difficult to see any impact as the number ranged from 72 days (due to the lag between the PDS cut and the rehearsal date) to counts suggesting information had not been updated for more than 2000 years. We have since recalculated it to years and excluded any records showing an update more than 70 years ago. There are only a small number of records within each individual Reason for Removal, so we are considering combining these to a simple binary flag for whether someone has a Reason for Removal or not, regardless of what it is.

Evaluation

A proof of concept version of ACID was created from available data containing UPRN for the rehearsal areas at address level. Initial models have been applied to one of these areas. We therefore achieved between the medium and full viable product detailed in the introduction. Discussions are occurring within ONS about how ACID could be applied in the Census if a low response scenario occurred. Appendix 6 lists the assumptions that were made in the creation of the proof of concept ACID.

Producing the proof of concept ACID has taught us valuable lessons about the admin data sources and processes which would be used for a full ACID. This work benefitted from strong collaboration between areas of ONS.

Since ACID is address-centric it is necessary for all data sources to include UPRN to allow data linkage. Where address information is available in the admin data ONS are able to apply a UPRN. However not all admin data sources received by ONS include this information. This therefore limited the number of datasets which could be used to create the proof of concept.

ONS is in discussion with data suppliers to request access to additional datasets which provide evidence of activity at addresses or enable us to place individuals within addresses.

Future Steps

The main future step for this project is to expand the ACID data frame. So far, this project has looked at data that specifically relates to the date of the 2019 census rehearsal, or as close as possible. Ideally, we want to expand this project longitudinally by including datasets that cover the reference period of interest (2021 Census) but also include longitudinal data going back as far as possible/feasible. This would improve the confidence with which we can assign people to addresses and highlight continuity issues that can be removed. In the same vein, acquiring more datasets that contain UPRN would greatly improve the quality of our final table by improving the process by which we place individuals in households and flag signs of life.

ACID for use with the 2021 Census will need to be created for all the local authorities of England and Wales. Some of the datasets used in the proof of concept do not have full coverage across these. This can impact how the models are applied.

Different models will continue to be applied to the proof of concept version of ACID and analysed to determine how effective ACID has been at placing individuals in the correct households. This work will also consider different groupings of data by, for example, Local Authority as well as age amongst other factors.

It is beyond the scope of this paper to discuss timely availability and access to admin data as conversations are ongoing between ONS and data providers. ONS is also researching increasing the efficacy of the tool used to assign UPRN based on address information. A discovery is currently being undertaken which will hopefully improve accuracy and throughput and improve the process by which we geo-reference our data.

All linkage and analysis has occurred within ONS' Data Access Platform (DAP). A new project would be set-up and the ethics assessment revisited for a full version of ACID. This would be within the same area of DAP as the Census data being processed (a Role-Based Access System [RBAC]) which will improve the speed of access to data. Being included in the RBAC in 2021 would automatically increase integration with the parallel strands being worked on. Being part of the RBAC would also likely help with the dataflows in terms of downstream processing of ACID data. The pipeline of moving ACID work onwards from our team to another team has not been tested so far in this project but being inside the RBAC should facilitate this movement.

Work needs to be done to look further into the quality of all the admin datasets involved – as well as any that may be included in future. This will be particularly important if more refinement of the households used in the estimation is to be done through various modelling methods including comparisons of names.

A 'source of truth' for comparison for the model needs to be identified for England and Wales. The rehearsal data has been used for the proof of concept. We are exploring whether ONS survey data can be used for a full-scale ACID, which would act as a proxy for Census data.

We would appreciate if the Methodological Assurance Review Panel (MARP) would consider the following questions:

- Are there any challenges or risks to this approach that we haven't identified in the paper?
- Are there any suggestions for different modelling approaches?
- Are there any thoughts on managing the issue of different data availability in different areas?

References

ISC (2020). Independent Schools Council Research [online] Available at: https://www.isc.co.uk/research/ [Accessed 17 Nov. 2020]

ONS (2016). *The Census Coverage Survey* [online] Available at: https://www.ons.gov.uk/census/2001censusandearlier/designandconduct/theonenumbercensus/thecensuscoveragesurvey [Accessed 14 Oct. 2020]

ONS (2020a). *Admin-based population estimates and statistical uncertainty: July 2020* [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/adminbasedpopulationestimatesandstatisticaluncertainty/july2020 [Accessed 14 Oct. 2020]

ONS (2020b). *Statistical design for Census 2021, England and Wales* [online] Available at: https://www.ons.gov.uk/census/censustransformationprogramme/censusdesign/statisticaldesignforcensus2021englandandwales [Accessed 14 Oct. 2020].

Appendices

Appendix 1: ABPE vs. ACID

	ABPE	ACID
Aim of the dataset	Identify usual residents as a base population count from linked administrative datasets	Pre-populate the Census address frame with administrative records
Target population	Usual residents in England and Wales.	All residents listed in residential addresses in England and Wales
Inclusion rules	Rules developed to remove all instances of over-coverage, at the expense of residual under-coverage that can be estimated for using a survey	There are no rules for removing records or addresses. Instead likelihood indicators are given for address occupation and individuals being resident.
Location assignment	Rules assigning to location are secondary to determining usual residence in the country. Some location assignments are at postcode level if addresses are not available.	Assignment is always at address level (UPRN) and are based on PDS (Personal Demographic Service) data. All other data sources with location information are used to validate/invalidate PDS as having the correct address information.
Estimation method	Primary designed for dual system estimation (DSE) framework.	Primarily designed for weighting class framework (estimating populations of households not responding to 2021 Census).
Availability	A dataset created for internal use by ONS. Information learned in the creation will feed into the creation of the population statistics to be published.	Created for publication as well as ONS use.

Appendix 2: Match Keys

- Match Key (MK) 1
  - Full Name, Date of Birth, Sex, UPRN
- MK 2
  - Forename, Surname, Date of Birth, Sex, UPRN
- MK 3
  - Levenshtein Full Name, Date of Birth, Sex UPRN
- MK 4
  - Soundex Forename, Soundex Surname, Date of Birth, Sex, UPRN
- MK 5
  - Forename, Surname, Date of Birth, UPRN
- MK 6
  - Levenshtein Forename, Levenshtein Surname, Date of Birth, Sex, UPRN
- MK 7
  - Levenshtein Forename-Surname, Levenshtein Surname-Forename, Date of Birth, Sex, UPRN
- MK 8
  - Middle Name, Surname, Date of Birth, Sex, UPRN
- MK 9
  - Levenshtein Forename, Levenshtein Middle Name, Date of Birth, Sex, UPRN
- MK 10
  - Forename-Middle Name, Surname, Date of Birth, Sex, UPRN
- MK 11
  - Middle Name-Forename, Surname, Date of Birth, Sex, UPRN
- MK 12
  - Forename, Date of Birth, Sex, UPRN
- MK 13
  - Levenshtein Forename, Date of Birth, Sex, UPRN
- MK 14
  - Bigrams Forename, Surname, Date of Birth, Sex, UPRN
- MK 15
  - Forename, Surname, Age, Sex, UPRN
- MK 16
  - Forename, Surname, Day of Birth, Month of Birth, Year of Birth (within 10), Sex, UPRN
- MK 17
  - Forename, Surname, Year of Birth, Sex, UPRN
- MK 18
  - Levenshtein Forename, Levenshtein Surname, Levenshtein Date of Birth, Sex, UPRN
- MK 19
  - Levenshtein Forename, Levenshtein Surname, Age (within 2 years), Sex, UPRN
- MK 20
  - Levenshtein Forename, Levenshtein Surname, Date of Birth, Sex (agree or missing), UPRN
- MK 21
  - Levenshtein Forename, Levenshtein Surname, Age, Sex (agree of missing), UPRN

Best matches by match key

MATCH KEY	COUNT	%
1	130,467	70.66
2	37,529	20.32
3	4,290	2.32
4	2,069	1.12
5	512	0.28
6	862	0.47
7	163	0.09
8	0	0
9	924	0.50
10	903	0.49
11	421	0.23
12	1,855	1.00
13	500	0.27
14	1,589	0.86
15	1,761	0.95
16	417	0.23
17	99	0.05
18	144	0.08
19	119	0.06
20	16	0.01
21	13	0.01
TOTAL	184,653	100.00

Appendix 3: Modelling Variables

Outcome	More Information	Source
match	Signifies whether or not a match has been found between the PDS data and the Census Rehearsal Responses. Linkage only done for UPRNs for which we received a Rehearsal response	PDS-Rehearsal linakge
Covariates	**More Information**	**Source**
rfr_xyz	This information comes from the Reasons for Removal variable in the PDS. This indicates if either the patient or practice has requested they be removed from the data and why. It has been split into 16 separate columns for ease of modelling which are: - AFN: Armed Forces enlistment, notified by Armed Forces - CAN: Cancelled - CGA: Gone away – address not known/FP69 - DEA: Death - EMB: Embarkation - NIT: Transferred to Northern Ireland - OPA: Address out of practice area - ORR: Other reason - RDI: Practice request immediate removal - RDR: Practice request removal - RFI: Removal from residential institute - RPR: Patient request removal - SCT: Transferred to Scotland - SDL: Services dependent, notified locally - TRA: Temporary resident not returned - X: No current NHAIS posting	PDS
ctax_ls	Maximum surname similarity score from comparing every surname associated with a UPRN between PDS and Council Tax. Based on Levenshtein edit distance.	PDS-Council Tax comparison
er_ls	Maximum surname similarity score from comparing every surname associated with a UPRN between PDS and Electoral Register. Based on Levenshtein edit distance.	PDS-Electoral Register comparison
esc_ls	Maximum surname similarity score from comparing every surname associated with a UPRN between PDS and English School Census. Based on Levenshtein edit distance.	PDS-English School Census comparison
esc_age_match	Flag to identify if both the PDS and English School Census have a child of the same age associated with the UPRN	PDS-English School Census comparison
xyz_years	Number of years between the various date variables in the PDS and the Census Rehearsal date. These dates are: - Business_effective_from_date - Nhais_posting_bef_date - Rfr_bef_date - Address_bef_date	PDS

Appendix 4: Logistic Regression Detail

Logistic Regression Summary		Full Logistic Regression Model		Reduced Logistic Regression Model	
		Odds Ratio	Significance	Odds Ratio	Significance
Electoral Register Surname Similarity		3.93	0.000	4.24	0.000
Council Tax Surname Similarity		3.26	0.000	3.63	0.000
Business Effective from Date difference - years		0.999	0.000	0.999	0.000
Address Business Effective from Date difference - years		0.974	0.000	0.960	0.000
Reason for Removal	**Practice request removal**	0.776	0.781	-	-
	Address out of practice area	0.0523	0.598	-	-
	Other reason	0.00954	0.000	0.00908	0.000
	Transferred to Scotland	0.00526	0.000	0.00662	0.000
	Gone away – address not known	0.00237	0.000	0.00255	0.000
	Embarked (emigrated)	0.00231	0.000	0.00260	0.000
	Transferred to Northern Ireland	0.000	0.993	-	-
	Cancelled	0.000	0.987	-	-
	No current NHAIS posting	0.000	0.998	-	-
	Death	0.000	0.999	-	-

Appendix 5: Random Forest Model Detail

Random Forest Modelling Summary		Feature Importance	
		Full Random Forest Model	Reduced Random Forest Model
Electoral Register Surname Similarity		0.359	0.423
Council Tax Surname Similarity		0.242	0.265
Business Effective from Date difference - years		0.142	0.160
Address Business Effective from Date difference - years		0.142	0.152
Reason for Removal	**Cancelled**	0.0358	-
	Gone away – address not known	0.0192	-
	Embarked (emigrated)	0.0186	-
	Transferred to Scotland	0.0130	-
	Death	0.0130	-
	Other reason	0.0126	-
	Transferred to Northern Ireland	0.00259	-
	No current NHAIS posting	0.000388	-
	Practice requested removal	0.000121	-
	Address out of practice area	0.000	-

Appendix 6: Assumptions and Limitations

- Rehearsal Frame being a complete record of all residential addresses
    - CEs weren't included in 2019 rehearsal
    - Rehearsal Frame wasn't updated in field during 2019 rehearsal
- Communal Establishments were removed from the Rehearsal Address frame before starting any analysis
- Special populations not included in rehearsal data
    - These are being left out of the proof of concept work. Ideally, we would include aggregate counts of special populations from Ministry of Defence, Ministry of Justice etc.
- Timeliness of data
    - Ideally, we would need admin datasets that relate to Census day and are therefore extracted as close to Census day as possible. If this is not possible then data extract dates should be as close to one another as possible.
    - For this work, we have used English School Census data from January 2019, Personal Demographic Service from August 2019 and the rehearsal responses from October 2019
- Assumption of individuals within households who share a surname being related
    - Also introduces possibility of error with, for example, blended families, single parent families, etc.
- Assumption that Census Rehearsal data is perfect
    - Has zero within household non-response
- Assumption that rehearsal data is representative
    - We only received around 30% response rate and we are looking to apply knowledge gained from this to the rest of the data
    - In a 'real' application, even a very low response rate is likely to be significantly higher than this
    - Also assume that the relationships between the admin data profiles and the likelihood of being in the correct address are the same for the non-responding households as they are for the responding ones
- Do not count those who died or were born on rehearsal day (1351 deaths nationwide, 8 in rehearsal areas on rehearsal day)