Second Meeting of the
National Statistician's Data Ethics Advisory Committee

————————

Agenda and Papers

————————

Wednesday 14 October 2015

10:30 – 14:30

Board Room, UK Statistics Authority
Drummond Gate, London

National Statistician's Data Ethics Advisory Committee

Minute

Wednesday, 14 October 2015
Boardroom, Drummond Gate, London

Present
Board Members
Mr Ian Cope (Chair)
Mr Robert Bumpstead (Deputy Chair)
Mr Colin Godbold
Ms Annie Hitchman
Ms Isabel Nisbet
Ms Marion Oswald
Professor Martin Severs
Dr Dean Machin

UK Statistics Authority
Mr Adil Deedat
Dr Simon Whitworth

Office for National Statistics
Mr Owen Abbot (for item 8)

Apologies:
Mr Neil McIvor
Mr Osama Rahman
Mr Hetan Shah

1. Minutes and matters arising from the previous meeting

1.1 The Chair informed the committee that the minute of the first meeting had been agreed and signed off by correspondence. The minute, agenda and papers from the first meeting are now published on the UK Statistics Authority Website.

1.2 The Chair also welcomed Professor Martin Severs and Dr Dean Machin to their first meeting of the Committee.

1.3 Progress with actions from the previous meeting held on 6 July was reviewed. All actions were complete or on the agenda for further discussions.

2. Chair's report

2.1 The Chair asked members for feedback on a training session which was organised by the Secretariat. The sessions, which covered legislation and data sharing, administrative data linkage and statistical disclosure control were well received by members who attended.

2.2 The Chair presented details of datasets, held by other government departments and agencies, which the Office for National Statistics is currently looking to acquire and what statistical purposes they would be used for. The meeting heard that requests for these datasets would be made using appropriate legal gateways (including information sharing orders).

2.3 NSDEC members sought more information as to what an address register was and how it would be used. The meeting heard that there was no complete address list in existence in England and Wales, which would be fit for purpose for a Census. In the last Census a combination of sources such as the postcode address file and the national address gazetteer were used to produce a full list of residential, communal and business addresses. This was needed to ensure Census forms were sent to all relevant household and communal establishment addresses in England and Wales. In addition, a manual field check was carried out by ONS to check whether there had been any changes to addresses, for example a house being converted into flats.

2.4 It was agreed that an item on the address register would be presented at a future meeting.

3. Ethical Principles [NSDEC(15)04]

3.1 Mr Bumpstead introduced a revised draft set of ethical principles and informed NSDEC members that their comments from the first meeting, in addition to those received by correspondence, had been integrated into the most recent version.

3.2 It was suggested that the National Statistician may wish to have the principles endorsed by the Authority Board and instilled as ethical principles for the Office for National Statistics and the Authority more broadly.

3.3 Given NSDEC's advisory role, it was suggested that the use of the word 'ensure' in the ethical principles may be too strong a requirement. This could be amended by removing references to ensure and leaving principles as single statements.

3.4 It was agreed that principle four should have a reference to the common law duty of confidence included.

3.5 The meeting agreed that the ethical principles would be amended to reflect member's comments and would be circulated to members for sign off ahead of the next meeting.

ACTION: Secretariat to redraft ethical principles and send to members.

4. Terms of Reference [NSDEC(15)05]

4.1 Mr Bumpstead introduced the revised version of the terms of reference which had been redrafted following comments from members at the first meeting and subsequently by correspondence.

4.2 Consistent with comments received for the ethical principles, members advised that the word ensure should be replaced with advise.

4.3 The meeting agreed that the terms of reference would be amended to reflect member's comments and would be circulated for approval before the next meeting.

ACTION: Secretariat to redraft terms of reference and send to members.

5. Word from the National Statistician

5.1 The Chair introduced and welcomed Mr John Pullinger, National Statistician.

5.2 Mr Pullinger outlined priorities for the statistical system over the next five years. These included improving economic statistics, mobilising data to make better decisions and improving capability to ensure that the statistical service have the tools and skills to use data for maximum impact. Mr Pullinger stated that NSDEC had an important role to play in setting the ground rules for the ethical discussions of these priorities and that the Committee's remit reflected the span of his research and statistics responsibilities. The following points were made in discussion:

i. it was agreed that the National Statistician would seek the endorsement of the Authority Board for the ethical principles;
ii. the Committee felt that it may be useful to consider the appropriate redress to be given to members of the public if their personal information was disclosed or misused; and
iii. the work of the Committee needs to be engaged with, and informed by, wider work on public acceptability.

ACTION: Secretariat to liaise with Authority Board Secretariat for the ethical principles to be considered for endorsement by the Authority.

6. Application: process and policies [NSDEC(15)06]

6.1 Mr Deedat provided an overview of what an application process could look like for a researcher applying for ethical review from NSDEC. The meeting was informed that the relevant ethical frameworks had been used to guide the suggested operational practices for NSDEC.

6.2 The meeting felt that the role of NSDEC would be to advise the National Statistician. It was therefore thought that any appeals would be against the National Statistician's decision rather than NSDEC.

6.3 It was also agreed that the role of NSDEC would be different depending on where the projects originated from:

i. advice given to projects originating from ONS would be definitive unless over ruled by the National Statistician who, as the Chief Executive of the Authority and ONS, is accountable for their work. Therefore, these projects would not require an appeals process;
ii. for projects from the Government Statistical Service, any outcome would be advisory to the department concerned and therefore no appeals process would be required; and
iii. for ADRN projects, the Committee would provide ethical advice on projects from the Government and third sector research communities for the ADRN Approvals Panel. The meeting felt that an appeals panel was required for these projects. It was suggested that the ADRN Appeals Panel, which is a subcommittee of the ADRN Board could also act as a group to whom researchers could appeal to.

6.4 The Committee agreed in the first instance that the time taken to provide an outcome on a project to the researcher along with written feedback would be up to sixty working days following submission of a project.

6.5 It was suggested that when projects are completed that an abstract relating to the research is published.

ACTION: Secretariat to amend the paper to outline the role of NSDEC for different projects and to discuss an additional role for the ADRN Appeals Panel with the Chair of the ADRN Board.

7. Project guidance and template [NSDEC(15)07]

7.1 Mr Deedat introduced the guidance and template which researchers would use in order to apply for an ethical review. The Committee heard that these had been developed through a review of application forms from other ethics committees and through consultation with relevant parties to learn from their experience and expertise.

7.2 It was suggested that the sections in the guidance relating to adverse events be combined. Members also agreed the need to ensure an appropriate mechanism for escalating an adverse event using the recommendations of the Information Commissioner's Office.

7.3 Members suggested that the ethical principles be included in the guidance where applicants are asked to consider the ethical implications of their research.

ACTION: Secretariat to amend the guidance and the template to reflect member's comments.

8. Estimating ethnicity from Names [NSDEC(15)08]

8.1 The Chair declared that he is the Information Asset Owner for Census data, and would be taking decisions on this project in light of the committee's advice. He therefore passed chairing responsibilities for this item to the deputy Chair, Mr Bumpstead.

8.2 Mr Abbott, senior methodologist in ONS, presented the research proposals. The meeting heard that, following comments during the first meeting, the ONAMAP tool, which attempts to derive an ethnicity from a name, had been removed by UCL from the internet.

8.3 It was reported that changes to the project had been made following initial feedback from the first meeting. This included reporting probabilities of a name being associated with ethnicities rather than a single ethnicity being outputted.

8.4 The Committee made the following comments and suggestions:

i. if ONS were to offer their data and services to help improve the quality of the tool for University College London (UCL), then they would also need to provide a similar service should other organisations with a similar tool approach them;
ii. more information is required detailing who would be permitted to use the tool, for what purposes, and under what arrangements; and

iii. some consultation with groups representing different ethnic groups was required, to understand acceptability issues.

8.5 Members agreed that the project should be classed as requiring major revisions and should be further considered by the Committee. At the next meeting further clarification should be given regarding:

i. the intellectual property rights for the tool and who, following further development, would be able to make use of it;
ii. the measures UCL and users of the tools have taken to ensure they have satisfied the requirements of the Data Protection Act around processing of sensitive personal data and other legal requirements; and
iii. ongoing use of the tool.

8.6 The Committee agreed that notwithstanding the need for further information, the project could provide significant benefits in this area and were minded to advise approving the project if concerns can be adequately addressed.

ACTION: Mr Abbott to provide additional information requested by NSDEC.

9. Review of ONS release practices

9.1 Dr Whitworth provided a presentation which informed members of a review which had been commissioned by the Chair of NSDEC. The review considered the practices by which ONS release securely anonymised data and made a number of recommendations.

9.2 Dr Whitworth outlined the recommendations which had been made to improve ONS release practices. These covered governance, transparency and reporting with regards to release practices.

9.3 The meeting was informed that the item would be brought back to the Committee for further consideration once a plan of action had been agreed by responsible business areas. It was also reported that the Chair of NSDEC would be commissioning a complementary review of the governance arrangements of data that was flowing into the office and that this would be discussed at a future meeting.

10. Any other business

10.1 A future meeting of NSDEC may take place at the Titchfield site. Included within this visit would be an opportunity for members to visit relevant business areas such as the big data lab.

UK STATISTICS AUTHORITY

NATIONAL STATISTICIAN'S DATA ETHICS ADVISORY COMMITTEE

Agenda

Wednesday, 14 October 2015
Board Room, One Drummond Gate, London
10:30am – 2:30pm (coffee from 10:00am)

Chair: **Mr Ian Cope**

Apologies: **Mr Neil McIvor**
Mr Osama Rahman
Mr Hetan Shah

(10:30am to 12:00pm)

1 10:30am	Minutes and Matters arising from the previous meeting	Mr Ian Cope
2 10:40am	Chair's report Acquisition of new data sources	Oral Report Mr Ian Cope
3 10:55am	Ethical principles	NSDEC(15)04 Mr Robert Bumpstead
4 11.20am	Terms of reference	NSDEC(15)05 Mr Robert Bumpstead
5 11:30am	Word from the National Statistician	Oral Report Mr John Pullinger

Lunch (12:00pm to 12:30pm)

(12:30pm to 2:30 pm)

6 12:30pm	Applications: policies and process	NSDEC(15)06 Mr Adil Deedat
7 1:00pm	Project template and guidance	NSDEC(15)07 Mr Adil Deedat
8 1:30pm	Estimating ethnicity from names	NSDEC(15)08 Mr Owen Abbott
9 2:00pm	Review of ONS release practices	Oral Presentation Dr Simon Whitworth
10 2:20pm	Any other business	

Next meeting: Wednesday 27 January 2016

National Statistician's Data Ethics Advisory Committee

Minute

Monday, 6 July 2015
Boardroom, Drummond Gate, London

Present
Board Members
Mr Ian Cope (Chair)
Mr Robert Bumpstead (Deputy Chair)
Mr Colin Godbold
Ms Annie Hitchman
Mr Neil McIvor
Ms Isabel Nisbet
Ms Marion Oswald
Mr Osama Rahman

Also in Attendance
Mr Adil Deedat
Dr Simon Whitworth
Ms Jane Naylor for item 5
Mr Peter Stokes for item 7
Ms Lucy Vickers for Item 7

Apologies
Professor Martin Severs
Mr Hetan Shah

1. Minutes and Matters arising from the previous meeting

1.1 The minute of the meeting on 6 July 2015 was agreed by correspondence and signed off by the Chair.

2

Chair’s report

Mr Ian Cope

List of Annexes

Annex A Acquisition of new data sources, Adil Deedat, NSDEC Secretariat, Central Policy Secretariat, 7 October 2015

Annex A – Acquisition of new data sources

Data Set	Information
PAYE and Benefits information	i. Contains record level information which includes Encrypted National Insurance Number (NINo) along with benefits and annual total PAYE and tax credit information. ii. This will then be linkable to existing DWP data within ONS using encrypted NINo. iii. Could be used to produce income estimates and enhanced Census outputs. iv. Data expected Autumn 2015.
All England Education Dataset	i. Created by the department for Business Innovation Skills using multiple data sources. ii. Contains records back to 1984 and include information relating to educational attainments (GCSEs, A-levels, degrees). iii. Data expected Spring 2016.
Health Demographic	i. Early stages of acquisition. ii. Dataset would indicate whether an address on the GP Patient Register has been verified with the individual when they visited a health professional. iii. Could help assess the quality of address information on other administrative sources and accuracy of population estimates. iv. Data expected late 2016.
Driver and Vehicle Licensing Agency data	• Acquisition work beginning. • DVLA data which could provide information on car ownership which is collected in the Census.
TV Licensing	• Acquisition work beginning. • Could provide information for an address register.
Armed forces	• Acquisition work beginning. • Record level information on armed forces for use in population estimates.

3

Ethical Principles

These have now been published on NSDEC’s web pages

4

Terms of Reference

These have now been published on NSDEC’s web pages

5

Word from the National Statistician

Mr John Pullinger

UK Statistics Authority

National Statistician's Data Ethics Advisory Committee

NSDEC(15)06

Applications: policies and process

6

Purpose

1. This paper outlines the policies and process by which researchers will apply for ethical review.

Recommendations

2. Members of NSDEC are invited endorse the policies and processes proposed.

Background

3. NSDEC has a UK wide remit and will provide advice on research projects originating from the Office for National Statistics (ONS), the government statistical service (GSS) and the Administrative Data Research Network (ADRN).

4. NSDEC is scheduled to meet four times a year. Whilst this is sufficient for the current volume of projects, as NSDEC becomes more established, demand is likely to increase. It is envisaged that this increase will be, in part, due to an increase in ADRN projects originating from government and the third sector.

5. The ADRN Approvals Panel is tasked with ensuring that projects are legal, of scientific merit and of public or policy benefit. The Approvals Panel must also satisfy themselves that the project they are considering has undergone and passed an ethical review.

6. As the Approvals Panel meets monthly, NSDEC may require a mechanism by which they can provide ethical review of ADRN projects between NSDEC meetings so as not to stall researcher's projects from progressing.

7. Ethics committees have good practice operating procedures. These include defining the possible outcomes of ethical review, the time by which an applicant can expect a response from a review, and a process by which the applicant can appeal against the outcome of the ethical review of their project.

8. Here we outline how NSDEC could review projects in the early phase of its operation and present a future process, should demand for ethical review increase. We also present a number of polices to ensure that applicants' projects can be progressed in as timely and as fairly a manner as possible.

Discussion

Initial process for ethics review

9. In early stages of operation it is probable that demand from ADRN and GSS projects are likely to be low, and therefore reviewing projects at quarterly meetings should be sufficient.

10. In this start up phase researchers applying for ethics review will download and complete the project template **[NSDEC(15)07]** from the UK Statistics Authority website and submit the application to the Secretariat.

11. The Secretariat will check the application to ensure that it is valid; plain English is used and all mandatory fields are completed satisfactorily.

12. The Secretariat will then provide the application with a cover note in the Board pack ahead of the next meeting.

13. At the next available meeting, NSDEC will discuss the project and come to a consensus on how to proceed. In line with other relevant ethics committee, the options for the committee would be to advise the National Statistician to:

    i. approve the research and allow it to proceed;
    ii. approve the research subject to minor revisions. These could be checked for compliance by the Secretariat;
    iii. recommend major revisions to the research. The researcher would need to reconsider the research in light of recommendations made by the major revisions and return the application to a future committee meeting; and
    iv. reject the research advising that the research be stopped from proceeding.

14. The role of NSDEC and therefore the National Statistician differs depending on where the research proposal originates from.

    i. **Office for National Statistics:** any outcome would be definitive unless overruled by the National Statistician, who as Chief Executive of the Authority and ONS is accountable for their work. In this instance there is no appeals mechanism.

    ii. **Government Statistical Service:** any outcome would be advisory to the department(s) concerned and therefore no appeals process would be required.

    iii. **Administrative Data Research Network:** NSDEC would provide advice to the ADRN Approvals Panel on projects from the government and third sector research communities. Appeals relating to ADRN projects could be directed to the ADRN's Appeals Panel.

15. Deliberations on projects will be given at the meeting and the Secretariat would be responsible for ensuring that the applicant is contacted with the outcome and feedback from the Committee.

16. Researchers applying to NSDEC for ethical review will receive written feedback and the outcome of the review within sixty working days from date of submission.

17. The title of the project and a summary along with NSDEC's decision will be published on the UK Statistics Authority website once the committee have deliberated on a project and provided feedback and the outcome to the applicant.

Future process of ethics review

18. As the demand for ethical review increases over time, NSDEC may need a mechanism to ensure that projects are deliberated on in periods in between meetings of NSDEC.

19. The process by which researchers apply for ethical review and the outcomes provided are similar to those outlined above (in the initial process for ethics review), with two notable differences:

    i. a subcommittee of NSDEC would convene either by correspondence or face to face meetings to review projects in-between meetings of NSDEC; and

    ii. where applicants feel that due process has not been followed, they may appeal to NSDEC.

20. As with the initial application process and policies, the Secretariat would provide the outcome and feedback on the application within a predefined time period.

Adil Deedat, NSDEC Secretariat, Central Policy Secretariat, 1 October 2015

UK Statistics Authority

National Statistician's Data Ethics Advisory Committee

NSDEC(15)07

Draft Project Template and Guidance

Purpose

1. This paper proposes a template and guidance for use when applying to the National Statistician's Data Ethics Advisory Committee (NSDEC) for ethical review.

Recommendations

2. Members of NSDEC are invited to consider and comment on:
   i. the draft guidance **(Annex A)**; and
   ii. the draft template **(Annex B)**.

Background

3. At the first meeting of NSDEC on 6 July 2015, the need for a standardised template for researchers to use when applying for ethical review of research projects was agreed.
4. As NSDEC has an UK wide remit it can offer its advice on policies and projects involving the access, use and sharing of data from the Office for National Statistics (ONS), the Government Statistical Service (GSS) and the devolved administrations.
5. NSDEC will also carry out an ethical approval function for Administrative Data Research Network (ADRN) projects originating from government and the third sector. The UK Statistics Authority provides the governance for the ADRN.

Discussion

Templates and guidance

6. The template and guidance have been developed through considering other existing ethics committee review forms and through engagement with relevant parties within the Administrative Data Service, the service which supports the ADRN.
7. A large number of university ethics application forms from within academia offer similar ways of reporting project details and information. Most offer the researcher an opportunity to summarise the research in plain English, to consider the ethical implications of their work and to identify benefits.
8. In many of these application forms the main focus is on primary data collection, for example undertaking surveys or collecting biological samples from individuals. Whilst robust for the needs of primary data collection and related research, there is less detail around secondary data analysis.
9. The ADRN Approvals Panel is responsible for ensuring ADRN project proposals are of scientific merit, legal and of public or policy benefit. Whilst the Panel do not provide ethical scrutiny, they must assure themselves that an ethical review has been conducted.
10. Further to issues experienced by the ADRN Approvals Panel, the Administrative Data Service has subsequently issued guidance to academic institutions on ethics around data linkage to ensure that research involving administrative data linkage is considered appropriately and not solely as secondary data analysis.

11. In considering the merits of different templates as well as the guidance from the Administrative Data Service, a template has been developed to ensure that all data relevant themes are captured.
12. The template and guidance for use have received positive feedback from a number of ONS business areas and has been reviewed by ONS's social surveys division with their recommendations implemented to ensure clarity.
13. The UK Statistics Authority's good practice team has also reviewed the template and guidance. They have recommended that the guidance be incorporated in to the template. This is currently being explored.
14. The guidance and template have been reviewed by relevant teams within the Office for National Statistics and has been piloted by the research proposal presented in **[NSDEC(15)08]**.

Adil Deedat, NSDEC Secretariat, Central Policy Secretariat, 17 September 2015

List of Annexes

Annex A Guidance for NSDEC project applications

Annex B Project Application forms

Annex A Guidance for Applying to NSDEC

This has now been published on NSDEC's web pages

Annex B Template to apply for NSDEC review

This has now been published on NSDEC’s web pages

UK STATISTICS AUTHORITY

National Statistician's Data Ethics Advisory Committee

NSDEC(15)08

Estimating ethnicity from names

Purpose

1. This paper presents a project proposal for the Office for National Statistics (ONS) to measure the quality of a tool, created by University College London (UCL), which estimates ethnicity from names. The proposal would also see ONS contribute to further develop the tool.

Recommendations

2. Members of NSDEC are invited to consider the project proposal at **Annex A** and advise the National Statistician to:

   i. approve the research and allow it to proceed;
   ii. approve the research subject to minor revisions;
   iii. recommend major revisions to the research and request the proposal be resubmitted to a future meeting once implemented; or
   iv. reject the research advising that the research be stopped from proceeding.

Background

3. ONS is exploring ways to produce new or update existing outputs from data such as those held for administrative purposes. Often, administrative sources contain names but do not contain important characteristics such as ethnic group.

4. UCL has developed a tool which allows users to estimate the ethnicity distribution of a population for which names (forename/surname combinations) are available.The existing tool was built using data that either did not include the whole population (e.g. the public version of the Electoral Roll) or were unlikely to fully represent the UK population (e.g. consumer data). These sources do not include individuals' self-assignments of their ethnic groups, making the classification more remote from the population that is being classified than is desirable.

5. The project aims to explore how 2011 Census data, which included names, self classified ethnicity and language spoken, can be used to understand and improve the accuracy of the estimates produced by the tool.

6. Following this work, a new version of the tool will be made publically available, along with user guidance on how to interpret the estimates. This will enable users to apply the tool and understand its strengths and weaknesses, helping them to know where they can and cannot use it.

Simon Whitworth, Data Governance and Policy, UK Statistics Authority, 4 October 2015

List of Annexes

Annex A Application: estimating ethnicity from names, Owen Abbott, Methodology, Office for National Statistics

National Statistician's Data Ethics Advisory Committee

Application for Ethical Review

The Application Process

This is an application form for applying for ethical review from the National Statistician's Data Ethics Advisory Committee (NSDEC). You should use the additional guidance when completing this form.

The application form should be completed in ***plain English*** which is understandable to lay members and all abbreviations should be explained the first time they are used. The form should contain sufficient information to ensure a thorough ethical review can take place.

Please word process the form using Arial or Times New Roman font, size 11. Where necessary expand text boxes on the form to accommodate answers, but ensure word counts are adhered to where specified.

Where sections are not relevant to your study please mark as N/A.

On completion the responsible owner should sign the application form and send to: nsdec@statistics.gsi.gov.uk

UK Statistics Authority

Section A
Application Details

A1	Responsible Owner

Full Name: [redacted]	**Position:** Branch Head, Sample Design and Estimation Branch, Population Methodology and Statistical Infrastructure Division.
Address [redacted] ONS, Segensworth Road, Titchfield, PO155RR	**Email:** [redacted]
	Telephone: [redacted]
	Organisation: Office for National Statistics

Declaration to be signed by the responsible owner

I have met with and advised the applicant on the ethical aspects of this project design *(applicable only if the responsible owner is not the Applicant).*

I understand that it is a requirement for all researchers accessing the data to have undergone relevant training and to have either relevant security clearances or approved researcher status in order to access the data.

I am satisfied that the research complies with current professional, departmental and other relevant guidelines.

I will ensure that changes in approved research protocols are reported promptly and are not initiated without approval by the National Statistician's Data Ethics Advisory Committee.

I will provide notification when the study is complete if it or fails to start or is abandoned.

I will ensure that all adverse or unforeseen problems arising from the research are reported in a timely fashion to the National Statistician's Data Ethics Advisory Committee.

I will consider all advice received from the National Statistician's Data Ethics Advisory Committee and should I be unable to implement any of the recommendations made, I will provide reasoning in writing to the Committee.

Print Name: [redacted]
Signature:
Date: 6th October 2015

UK Statistics Authority

A2	Applicant Details (if applicant is not the responsible owner)

Full Name	Position
Address:	Email:
	Telephone:
	Organisation:

A3	Project Information

Project Title: Estimating ethnicity from names	
Start Date: March 2015	End Date: Estimated March 2016
Project Sponsor (select all that apply)	
☐ ONS ☑ Collaboration ☐ ADRN ☐ Other (please specify)	

A4	Collaboration and Sponsors

List of Collaborators/Sponsors	Details and relevant documentation relating to collaboration (you may attach copies of relevant documentation)
Office for National Statistics **University College London, Department of Geography**	This is a Joint project funded by ESRC under their secondary data analysis initiative phase 2 – 2013. The funding provided by ESRC is for UCL staff only. ONS and UCL have agreed and signed a MOU for this project. The MOU covers: data access, terms and conditions for data access, stakeholder engagement, publications clearance, breach and dispute procedures and project termination. **Previous ethical reviews** The use of the existing tool for NHS

	applications was reviewed by the Health Research Authority (HRA) who gave a favourable ethical review for a project titled "Small area, geodemographic profiling of health needs"[1]. The review highlighted specific patient benefits through application of this approach, for instance the identification of better medication for specific groups and increased participation in screening programmes.

A5	**Proposed Site of Research** (select all that apply)

Where will the research take place?

☑ ONS ☑ VML ☐ HMRC Data Lab

☐ ADRC-E ☐ ADRC-NI ☐ ADRC-W ☐ ADRC-S

☐ Other (please Specify)

Is this a secure site?

☑ Yes ☐ No

[1] See ethical approval record at http://www.hra.nhs.uk/news/research-summaries/small-area-profiles-of-health-needs/

UK Statistics Authority

Section B

Project Details

B1	**Please provide a brief high level summary of the research giving necessary background** ***(max 500 words)***

ONS is exploring ways to produce new or update existing outputs from data such as those held for administrative purposes, particularly where there is a strong user need. One method is to apply a statistical model or algorithm to predict a new variable from those already included on such sources.

ONS is considering the potential for using name data in these models. There are existing commercial tools which estimate ethnicity from data including forenames and surnames, however their quality and origin is unknown. Often, administrative sources contain names but do not contain important characteristics such as ethnic group. There is a strong user demand for ethnicity statistics both in their own right and also combined with other variables.

This proposal is for a joint project with academics to measure the quality of an existing tool which estimates ethnicity from name data, with a view to improving the tool based on the findings. The intention is to publish this as a free software tool, with accompanying metadata and user guidance so that users of the tool can obtain their own estimates together with quality measures.

The existing tool has been successful, in that it has been used by a range of health care and other organisations. However, it was built using data that either did not include the whole population (e.g. the public version of the Electoral Roll) or were unlikely to fully represent the UK population (e.g. consumer data). Crucially, such sources do not include individuals' self-assignments of their ethnic groups, making the classification more remote from the population that is being classified than is desirable. It has not used ONS data previously, other than in comparison with published area level census outputs. This has meant that the predictions can be poor for some population groups, for instance specific groups whose names have become anglicised (e.g. those of Caribbean ancestry) or those from groups that consider themselves assimilated into British society (e.g. bearers or Irish names). The extent of the uncertainty of the predictions using the current tool is unknown. In addition, the tool provides estimates which are a mixture of ethnicity, nationality and religion – having discussed with UCL the intention is to produce estimates only of self-assigned ethnicity and language spoken, avoiding the vaguer terms that are necessary in the current classification in the absence of Census data.

2011 Census data would be used to measure the quality of the estimates. The census provides self-classified ethnicity which is more aligned with user requirements for outputs, whereas the existing data for creating the tool is based entirely upon pairings of given and family names.

The estimation methodology is based on the use of clustering algorithms which group forenames and surnames into groups based upon observed pairings of given names and surnames. The resulting classification will be probabilistic in nature – for each name there

UK Statistics Authority

will be an associated probability that the individual belongs to any of a number of ethnic and linguistic groups.

B2	Data Use

Type of data	Data Level *Please specify the name of the data set*			
	Aggregate Data	**Identifiable Data**	**De-identified personal data**	**Anonymised/ pseudo anonymised**
Administrative data *(please specify, e.g. Patient Register 2011, School Census 2012 etc, in the relevant options adjacent)*				
Big Data *(please specify e.g. Twitter data, smart meters and mobile phones, in the relevant options adjacent)*				
Survey Data *(please specify e.g.LFS, BRES, etc in the relevant options adjacent)*				
Census Data *(please specify year, e.g. Census 2011 in the relevant options adjacent)*	UCL project members will only have access to aggregate level 2011 Census data in the VML.	ONS project members will use identifiable 2011 Census data to prepare aggregate level data.		
Other *(please specify e.g. Ordinance Survey Address register in the relevant options adjacent)*				

B3	How will information be kept confidential and data kept secure? *(max 500 words)*

Only security checked ONS staff have access to the individual level census data on a secure server. UCL will only have access to aggregate level diagnostics as detailed in the study

protocol.

ONS staff within methodology group will prepare and clean the 2011 Census unit level data within the CDME (Census Data Management Environment) environment. UCL will provide an algorithm which ONS will import into the CDME, run against the cleaned census data, produce agreed aggregate diagnostic tables and apply agreed thresholds and then submit these for export from the CDME. The census data custodian will check and approve the export from the CDME, whereupon the diagnostic tables will be transferred into the standard VML. The UCL researcher will then access the diagnostic tables through controlled access to the standard VML. UCL will then update and refine their algorithm, and pass the revised algorithm back to ONS to rerun and generate revised diagnostics and thus iterate around the process.

Only the UCL researchers named in the ESRC research proposal (Paul Longley, James Cheshire, Alex Singleton and Muhammad Adnan) and Kira Kowalska (a UCL Phd student) will work on this project. UCL researchers needing to access the diagnostics will have to be ONS approved researchers in order to access the VML (as per standard VML access protocols). The ONS methodology staff working on this project are Owen Abbott, Helen Ross and Adriana Castaldo.

The diagnostic tables will mostly consist of proportions (not counts) to minimise the perception of disclosure. In addition, thresholds will be applied to minimise disclosure risks (e.g. to ensure that very rare names do not appear in diagnostic outputs). For diagnostics that do not include forename or surname fields, a threshold of 3 will be applied. For diagnostics that include forename or surname fields, a threshold of 10 will be applied. These diagnostics are not identifiable. The data will be accessed in a controlled manner through the standard VML, which is being used as the mechanism to access this sensitive data for this project only.

B4	**Please provide details of the research protocol or methodology** ***(e.g. data linkage, web scraping etc) (max 500 words)***

Names are a valuable link to family history and to cultural heritage, and can thus have important applications to understanding migration and population structures. Since 2003, research led by Professor Paul Longley at University College London (UCL) has investigated the geographical concentrations of names and ethnicity in the British Isles, and devised methods to analyse this data for a range of applications.

This has resulted in a tool which allows users to estimate the ethnicity distribution of a population for which names (forename/surname combinations) are available. The tool has proved useful for organisations, particularly those in the health sector, to be able to add a predicted ethnicity class onto their patient database records where previously it was not available, or as a way of updating local profiles for service needs (e.g. for translators) as census data becomes out of date.

The project aims to explore how 2011 Census data, which included both names and self classified ethnicity and language spoken, can be used to understand and improve the accuracy of the estimates. This will enable a robust assessment of the existing tool's performance using the Census top level 18 ethnic groups, and suggest where improvements could be made. The project will tune and improve the existing tool. A new version will be made publically available, along with user guidance on how to interpret the estimates. This

will enable users to apply the tool and understand its strengths and weaknesses, helping them to know where they can and cannot use it.

Study protocol
ONS staff will apply the existing tool to a subset of 2011 Census data. These will include records with ethnicity, language spoken, religion, country of citizenship, area of the country, age (which bears a correspondence to the popularity of many forenames) and names. They will be used to produce aggregate diagnostics for UCL showing the success of their algorithm. Thresholds will be applied to these diagnostic tables, designed so that the frequency of a particular forename or surname is above a minimum number. This will be to ensure that no person can be uniquely identified within the tables, and therefore unique forenames or surnames will be excluded and not used in this project. Similar rules will be applied to cross tabulations. These procedures have been assessed by ONS disclosure control experts to ensure we maintain our commitment to the public on the confidentiality of their personal information. These aggregate, non-identifiable data will be made available to UCL through the VML, where the UCL staff will analyse the diagnostics and make improvements to their algorithm. The revised algorithm can then be passed back to ONS so that it can be re-applied to the secure microdata to derive new diagnostics. This process will be iterative. All UCL staff who have access to the non-identifiable data will have signed the Census Confidentiality Undertaking. No results or analysis will be allowed to leave the VML if they are disclosive or include any personal data that will identify an individual.

B5	**Please outline the proposed benefits of the project** ***(max 500 words)***

The main evidence for the benefit to the public comes from the application of the existing software.

The UCL software emerged from collaborations with two primary care trusts (PCTs) in ethnically diverse London boroughs to improve ethnicity designations in medical records, and targeted public health initiatives. For instance, a London borough used the software in a pilot project seeking to increase extremely low rates of breast cancer screening amongst women of African Caribbean descent. It used the names classification to identify the ethnic groups of women who missed screening, and then targeted resources and information accordingly, leading to an increase in the uptake of screening among African Caribbean women. This led to specific patient benefits through earlier identification of breast cancer for that specific group.

Other PCTs have used the tool to analyse GP referral patterns and admissions to accident and emergency facilities to measure equality of service usage.

Between 2008 and 2013, over 15 PCTs, local authorities and other government organisations have licensed the tool; for example, in 2011–2012 the Health Protection Agency (now Public Health England) used it in a survey of hepatitis and other blood borne viruses to explore whether transmission was related to ethnicity, while NHS Lothian licensed it in 2010 to assess access to public health services such as smoking cessation. Most of these public sector applications are for seeking improvements to services, ensuring equality of access to public services or assessing whether the quality of that service differs across ethnic groups.

UK Statistics Authority

However, the quality of the existing software is unknown. Therefore, the decisions made for these and future applications were in the absence of information about the accuracy of the outputs. Were this to be provided, then the decisions made would be better informed. In addition, if the underlying quality were to be improved by using additional information on language spoken, age, religion and citizenship then this would also provide additional benefits to service providers and the public.

Thus, for example: health researchers would be able to augment imprecise descriptors (e.g. 'African') that are appended to many health records in order to measure and monitor the effectiveness of (preventive and remedial) interventions across sub-populations; public attitude surveys could be used to identify the degree of cross community support for reassurance policing; universities would be able to establish the effects of their 'widening participation' initiatives in selection and recruitment policies in different subjects of study; and employers would be able to investigate the representativeness of their recruitment procedures with respect to their local labour markets. ONS plan to explore this sort of approach for providing improved ethnicity estimates either between censuses or in the absence of a census.

B6	**Please outline any ethical issues that might arise from the proposed study and how they will be addressed** ***(all research projects have some ethical considerations, so this section must not be left blank)***

a) There could be a perception that the tool uses census data as a 'lookup' for a name to provide an ethnicity. We would mitigate this by:

- ensuring that the tool always indicates the level of uncertainty for its estimates, e.g. through the provision of probabilities of a forename/surname combination belonging to a particular ethnic group (see section B6).
- Providing clear documentation on how the tool works (i.e. that it does not contain any census data)

b) There could be a risk that this tool is misused. For instance, someone with access to a list of names could use the tool to estimate ethnicity to discriminate against that group. This could happen anyway using the existing tool or existing commercial alternatives (or through a person guessing a persons ethnicity from names). This project will provide information about the accuracy of estimates which may reduce the likelihood of this happening, although it could not be prevented. One option we could explore would be to use a different licensing model, for instance users have to provide brief details before they can download the software.

c) There is a risk that the tool is labelled as 'ONS approved', and therefore it might be perceived to have a error free output despite the intended provision of transparent methodology and information about quality.

d) There is a risk that some of the categories in the classification may be ambiguous or inappropriate. This arises in the existing tool because it does not use any data on the groups that name bearers would assign themselves to. Use of census data will greatly reduce this risk because predictions will be modelled using information on the groups that individuals assign *themselves* to.

e) There is a risk that users will be insufficiently aware of the uncertainty that is associated with the estimates. This was apparent in a previous version of the Onomap website that allowed users to enter a single name and receive a single predicted ethnic group without any

associated level of uncertainty or possible alternatives. We have since discussed with UCL who have taken down the site, and we do not envisage using this approach in the future. In addition, all estimates will always be accompanied by an indication of uncertainty.

f) There are a number of existing tools marketed by commercial organisations. They do not necessarily provide information on the quality of the outputs and are not transparent in their methodology. There is a risk that users of such tools are currently making poor decisions for want of better alternatives and because the quality of estimates are not understood. This project will enable users free access to a tool and enable them to make better decisions through providing information on the quality of outputs and on the methods used to create the estimates.

g) The use of name data for anything other than data matching (or to produce primary outputs like baby names which are of public interest) is not something that ONS has undertaken before – although linking names to age is deemed acceptable elsewhere (see http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/), is modelled by commercial organisations in the UK and can be estimated for young people in the UK (excluding immigrants) using Birth Registration data.

h) There might be a perception that we are giving individual level census data to UCL. Preserving the confidentiality of personal information provided by the public on their census questionnaire remains ONSs utmost priority. The tool will not have any census data within it. Only security checked ONS staff have access to the individual level census data on a secure server. UCL will only have access to aggregate level diagnostics as detailed in the study protocol.

i) There is a risk that organisations use the tool to derive ethnicity instead of collecting it directly. For instance, if they have a statutory requirement to measure service provision under the Equalities Act. This could happen anyway using the existing tool. This project will provide information about quality which may reduce the likelihood of this happening, as there may be restrictions around the quality of such measurement, although it could not be prevented. In some cases, if the quality is good then this might provide a benefit in reducing burden on the public and costs.

B7	**How will the findings of the research be disseminated?**

Through research papers outlining the research methodology and findings – these will be authored by UCL but cleared by ONS.

Through the software being made available for free download (on a website, probably one hosted by UCL), together with metadata, user guidance and the above papers. The software will allow the user to import a list of forenames and surnames, and will return an output dataset which includes for each name the most likely ethnic group together with an estimated probability of that being correct, as shown below (using made-up data):

a) Input table

Forename	**Surname**
Wayne	Rooney
Didier	Drogba

UK Statistics Authority

Asmir	Begovic
Gareth	Bale
David	Silva
Sulzeer	Campbell
Shinji	Okazaki

b) Output table

Forename	**Surname**	**Estimated ethnic group**	**Estimated Probability that estimate is correct**
Wayne	Rooney	White: British	0.92
Didier	Drogba	Black: African	0.83
Asmir	Begovic	White: Other	0.89
Gareth	Bale	White: Welsh	0.85
David	Silva	White: Other	0.85
Sulzeer	Campbell	Black: British	0.75
Shinji	Okazaki	Asian: Other	0.91

8.1

If the research methodology allows, we would like the output to include the likelihood of that name belonging to all ethnic groups (with the corresponding probabilities that will obviously be small), so that the output might look like (assuming there are only really 5 ethnic groups for illustrative purposes and a '-' implies that a probability is less than 0.01):

c) Output table including all ethnicities

Forename	**Surname**	**Estimated probability of being**				
		White: British	White: Other	Black: British	Black: African	Asian: Other
Wayne	Rooney	0.92	0.05	0.02	0.01	-
Didier	Drogba	0.01	0.02	0.13	0.83	-
Asmir	Begovic	0.10	0.89	-	-	-
Gareth	Bale	0.14	0.85	-	-	-
David	Silva	0.07	0.85	0.03	0.02	0.03
Sulzeer	Campbell	0.02	0.05	0.75	0.17	-
Shinji	Okazaki	0.02	0.05	-	-	0.91

Section C

Details of Data Subjects

C1	**Data subjects to be studied**

Does the Study include all subsections of the population (i.e. all ages, sex, ethnic groups etc	☑ Yes ☐ No *If no please detail which subsections with justification(s) below*

Subsections of the population (including vulnerable groups) the project focuses on: N/A
Justification for focusing on these subsections or groups: N/A

C2	**Please detail consent given to use data specified in section B2**

The 2011 Census Information Asset Owner has provided consent to use the data through the VML access mechanisms, given that the data made available to UCL are aggregates and thresholds are applied to ensure so individual can be identified.

C3	**If you are using data held by a third party please detail how you will obtain this**

N/A

9

Review of ONS release practices

Dr Simon Whitworth

Any other business

10