



## Citizen Information Project

### Final Report: Annex 2: Stakeholder processes, systems and data

#### 2C: Stakeholder profiles

## Version Control

---

<b>Date of Issue</b>	14 <sup>th</sup> June 2005
<b>Version Number</b>	1.0

<b>Version</b>	<b>Date</b>	<b>Issued by</b>	<b>Status</b>
1.0	14/06/05	PJ Maycock / CC Sanger	Final report

## Metadata

---

Coverage	UK
Creator	Office for National Statistics, General Register Office, Citizen Information Project Team
Date Issued	13/6/05
Language	English
Publisher	Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ
Status	Approved by Project Manager
Subject	Data quality, sharing and processing
Subject.category	
Title	Citizen Information Project: Annex 2C: Stakeholder profiles

# Contents

---

<b>Stakeholder profiles</b> .....	<b>9</b>
<b>1. Preface</b> .....	<b>10</b>
<b>2. Related documents</b> .....	<b>10</b>
<b>3. Introduction</b> .....	<b>11</b>
<b>3.1 Scope of consultation</b> .....	11
<b>3.2 Tier 1 stakeholder interactions</b> .....	11
<b>4. Driver and Vehicle Licensing Agency</b> .....	<b>13</b>
<b>4.1 Business processes and events</b> .....	13
<b>4.2 Current systems</b> .....	14
<b>4.3 Initiatives</b> .....	14
<b>4.4 Current data quality: Questionnaire responses</b> .....	14
<b>4.5 Current data quality: Data trial results</b> .....	15
<b>4.6 Current data sharing</b> .....	17
<b>5. Department for Education and Skills</b> .....	<b>18</b>
<b>5.1 Business processes and events</b> .....	18
<b>5.2 Current systems</b> .....	18
<b>5.3 Initiatives</b> .....	18
<b>5.4 Current data quality</b> .....	19
<b>5.5 Current data sharing</b> .....	20
<b>6. Department for Work and Pensions</b> .....	<b>21</b>
<b>6.1 Business processes and events</b> .....	21
<b>6.2 Current Systems</b> .....	22
<b>6.3 Initiatives</b> .....	23
<b>6.4 Current data quality questionnaire</b> .....	23
<b>6.5 Current data sharing</b> .....	24
<b>7. e-Government Unit</b> .....	<b>25</b>
<b>7.1 Business processes and events</b> .....	25
<b>7.2 Current systems</b> .....	25
<b>7.3 Initiatives</b> .....	26
<b>7.4 Current data sharing</b> .....	26
<b>8. General Register Office (England, Wales and Scotland)</b> .....	<b>27</b>
<b>8.2 Business processes and events</b> .....	27
<b>8.3 Current systems</b> .....	27
<b>8.4 Initiatives</b> .....	27
<b>8.5 Current data quality: Data trial results</b> .....	28
<b>8.6 Current data sharing</b> .....	29

<b>9.</b>	<b><u>Immigration and Nationality Directorate</u></b> .....	<b>30</b>
9.1	<u>Business processes and events</u> .....	30
9.2	<u>Current systems</u> .....	30
9.3	<u>Initiatives</u> .....	30
9.4	<u>Current data sharing</u> .....	31
<b>10.</b>	<b><u>HM Revenue and Customs</u></b> .....	<b>32</b>
10.1	<u>Business processes and events</u> .....	32
10.2	<u>Current systems</u> .....	33
10.3	<u>Initiatives</u> .....	33
10.4	<u>Current data quality: Questionnaire response</u> .....	33
10.5	<u>Current data quality: Data trial results</u> .....	34
10.6	<u>Current data sharing</u> .....	35
<b>11.</b>	<b><u>National Health Service</u></b> .....	<b>36</b>
11.1	<u>Business processes and events</u> .....	36
11.2	<u>Current systems</u> .....	37
11.3	<u>Initiatives</u> .....	37
11.4	<u>Current data quality questionnaire</u> .....	38
11.5	<u>Current data sharing</u> .....	38
<b>12.</b>	<b><u>Office for National Statistics</u></b> .....	<b>39</b>
12.1	<u>Business processes and events</u> .....	39
12.2	<u>Current systems</u> .....	39
12.3	<u>Initiatives</u> .....	39
12.4	<u>Current data sharing</u> .....	40
<b>13.</b>	<b><u>UK Passport Service</u></b> .....	<b>41</b>
13.1	<u>Business processes and events</u> .....	41
13.2	<u>Current systems</u> .....	42
13.3	<u>Initiatives</u> .....	42
13.4	<u>Current data quality: Questionnaire response</u> .....	42
13.5	<u>Current data quality: Data trial results</u> .....	43
13.6	<u>Current data sharing</u> .....	44
	<b><u>Appendix A: Data assessment - DVLA</u></b> .....	<b>46</b>
<b>1.</b>	<b><u>Data structure</u></b> .....	<b>47</b>
<b>2.</b>	<b><u>Statistical summary</u></b> .....	<b>48</b>
<b>3.</b>	<b><u>Individual field analysis</u></b> .....	<b>50</b>
3.1	<u>Driver number</u> .....	50
3.2	<u>Date of birth</u> .....	50
3.3	<u>Gender</u> .....	52
3.4	<u>Place of birth</u> .....	53
3.5	<u>Name elements</u> .....	53

<b>3.6</b>	<u><a href="#">Address lines</a></u> .....	54
<b>3.7</b>	<u><a href="#">Account creation date</a></u> .....	55
<b>3.8</b>	<u><a href="#">Last update date</a></u> .....	56
<b>4.</b>	<b><u><a href="#">Identity duplication</a></u></b> .....	<b>61</b>
<b>4.1</b>	<u><a href="#">Date of birth, name and address matching</a></u> .....	61
<b><u><a href="#">Appendix B: Data assessment - GRO</a></u></b> .....		<b>63</b>
<b>1.</b>	<b><u><a href="#">Analysis of datasets</a></u></b> .....	<b>64</b>
<b>1.2</b>	<u><a href="#">Coverage</a></u> .....	64
<b>1.3</b>	<u><a href="#">Field analysis</a></u> .....	64
<b>2.</b>	<b><u><a href="#">Data structure</a></u></b> .....	<b>64</b>
<b>3.</b>	<b><u><a href="#">Statistical summary</a></u></b> .....	<b>65</b>
<b>4.</b>	<b><u><a href="#">Individual data item analysis</a></u></b> .....	<b>67</b>
<b>4.1</b>	<u><a href="#">Unique ID</a></u> .....	67
<b>4.2</b>	<u><a href="#">Reference number</a></u> .....	67
<b>4.3</b>	<u><a href="#">GRO births – Date of birth</a></u> .....	68
<b>4.4</b>	<u><a href="#">GRO deaths – Date of birth</a></u> .....	69
<b>4.5</b>	<u><a href="#">Verified date of birth</a></u> .....	71
<b>4.6</b>	<u><a href="#">Date of death</a></u> .....	72
<b>4.7</b>	<u><a href="#">Gender</a></u> .....	73
<b>4.8</b>	<u><a href="#">Place of birth</a></u> .....	73
<b>4.9</b>	<u><a href="#">Name elements</a></u> .....	74
<b>4.10</b>	<u><a href="#">Address elements</a></u> .....	76
<b>4.11</b>	<u><a href="#">Update and creation dates</a></u> .....	77
<b>4.12</b>	<u><a href="#">Miscellaneous data items</a></u> .....	80
<b>5.</b>	<b><u><a href="#">Identity duplication</a></u></b> .....	<b>82</b>
<b>5.1</b>	<u><a href="#">Date of birth, name and address matching</a></u> .....	82
<b>5.2</b>	<u><a href="#">Date of birth and name matching</a></u> .....	83
<b><u><a href="#">Appendix C: Data assessment - GRO(S)</a></u></b> .....		<b>84</b>
<b>1.</b>	<b><u><a href="#">Coverage</a></u></b> .....	<b>85</b>
<b>2.</b>	<b><u><a href="#">Field analysis</a></u></b> .....	<b>85</b>
<b>3.</b>	<b><u><a href="#">Identity duplication</a></u></b> .....	<b>85</b>
<b>4.</b>	<b><u><a href="#">Demographic analysis</a></u></b> .....	<b>86</b>
<b>5.</b>	<b><u><a href="#">Data structure</a></u></b> .....	<b>86</b>
<b>6.</b>	<b><u><a href="#">Statistical Summary</a></u></b> .....	<b>86</b>
<b>7.</b>	<b><u><a href="#">Individual Column Analysis</a></u></b> .....	<b>87</b>
<b>7.1</b>	<u><a href="#">Unique ID</a></u> .....	87
<b>7.2</b>	<u><a href="#">Date of birth</a></u> .....	88
<b>7.3</b>	<u><a href="#">Verified date of birth</a></u> .....	90

7.4	<a href="#">Date of death</a>	90
7.5	<a href="#">Gender</a>	91
7.6	<a href="#">Place of birth address fields</a>	92
7.7	<a href="#">Name fields</a>	93
7.8	<a href="#">Address fields</a>	94
7.9	<a href="#">Date of registration</a>	95
<b>8.</b>	<b><a href="#">Identity duplication</a></b>	<b>96</b>
8.2	<a href="#">Date of birth, name and address matching</a>	96
8.3	<a href="#">Date of birth and name matching</a>	97
<b>Appendix D: Data assessment - HMRC</b>		<b>98</b>
<b>1.</b>	<b><a href="#">HMRC data structure</a></b>	<b>99</b>
<b>2.</b>	<b><a href="#">Statistical summary</a></b>	<b>100</b>
<b>3.</b>	<b><a href="#">Individual column analysis</a></b>	<b>101</b>
3.1	<a href="#">National Insurance Number (Nino)</a>	101
3.2	<a href="#">Nino suffix</a>	102
3.3	<a href="#">Account status</a>	102
3.4	<a href="#">Date of birth</a>	103
3.5	<a href="#">Date of birth status</a>	105
3.6	<a href="#">Date of death</a>	105
3.7	<a href="#">Date of death status</a>	106
3.8	<a href="#">Gender</a>	106
3.9	<a href="#">Date of entry</a>	107
3.10	<a href="#">Date of registration</a>	108
3.11	<a href="#">Name sequence number</a>	109
3.12	<a href="#">Name elements</a>	110
3.13	<a href="#">Address elements</a>	112
<b>4.</b>	<b><a href="#">HMRC record currency analysis</a></b>	<b>116</b>
4.1	<a href="#">Scope and analysis of data</a>	116
4.2	<a href="#">Current records by demographic</a>	116
4.3	<a href="#">Change of address statistics</a>	117
4.4	<a href="#">Change of name statistics</a>	120
<b>5.</b>	<b><a href="#">Identity duplication</a></b>	<b>121</b>
5.1	<a href="#">Date of birth, name and address matching</a>	121
<b>Appendix E: Data assessment - UKPS</b>		<b>122</b>
<b>1.</b>	<b><a href="#">Analysis of dataset</a></b>	<b>123</b>
1.2	<a href="#">Coverage</a>	123
1.3	<a href="#">Field analysis</a>	123
1.4	<a href="#">Identity duplication</a>	123
1.5	<a href="#">Demographic analysis</a>	124

<b><u>2.</u></b>	<b><u>Data structure</u></b> .....	<b>124</b>
<b><u>3.</u></b>	<b><u>Statistical summary</u></b> .....	<b>125</b>
<b><u>4.</u></b>	<b><u>Individual column analysis</u></b> .....	<b>127</b>
<b><u>4.1</u></b>	<u>Passport number</u> .....	127
<b><u>4.2</u></b>	<u>Passport status</u> .....	127
<b><u>4.3</u></b>	<u>Date of birth</u> .....	127
<b><u>4.4</u></b>	<u>Miscellaneous fields</u> .....	130
<b><u>4.5</u></b>	<u>Gender</u> .....	130
<b><u>4.6</u></b>	<u>Place of Birth</u> .....	131
<b><u>4.7</u></b>	<u>Name elements</u> .....	131
<b><u>4.8</u></b>	<u>Address elements</u> .....	132
<b><u>4.9</u></b>	<u>Last Update Date</u> .....	134
<b><u>4.10</u></b>	<u>Account creation date</u> .....	135
<b><u>4.11</u></b>	<u>PASS document type</u> .....	136
<b><u>5.</u></b>	<b><u>Identity duplication</u></b> .....	<b>136</b>
<b><u>5.1</u></b>	<u>Date of birth, name and address matching</u> .....	136

## Stakeholder profiles

---

## 1. Preface

- 1.1.1 The Citizen Information Project Final Report recommends the creation of an adult population register that will deliver benefits by sharing basic contact information (name, address, date of birth etc) across the public sector. The report recommends that the development of a population register is implemented as part of the ID Cards Scheme by utilising the National Identity Register (NIR) and that in the interim a range of short term data sharing initiatives are explored further.
- 1.1.2 This annex provides an overview of key stakeholders' business processes, systems and data quality. This information was used to
- identify business opportunities for gaining benefits from a population register,
  - evaluate a range of technical 'straw men' options, which included options for utilising existing systems and data
  - understand the characteristics and assessing the quality of existing data as the basis for assessing the benefit to be derived from the proposed solutions – both short term data sharing and long term implementation of ID Cards National Identity Register as a population register.

## 2. Related documents

- 2.1.1 'Annex 2: Stakeholder processes, systems and data' comprises the following documents:
- Annex 2A: Overview
  - Annex 2B: Data quality framework
  - Annex 2C: Stakeholder profiles: This document
  - Annex 2D: Data trial comparative results
  - Annex 2E: Data trial comparative results: Appendices
  - Annex 2F: Current data sharing across government
  - Annex 2G: Other data quality initiatives

## 3. Introduction

### 3.1 Scope of consultation

- 3.1.1 Stakeholders have been particularly important in the CIP project. With the CIP vision of facilitating major economies, efficiencies and service improvements across the public sector, it has been vital to understand the requirements of stakeholders, their current systems and where it might be possible to effect improvements and thus obtain benefits.
- 3.1.2 During the Feasibility Study some limited assessment of stakeholder processes, systems and data was carried out. Within Project Definition more detailed discussions have taken place on a wider scale.
- 3.1.3 Stakeholders have been categorised into three groups within the project:
- **Tier 1:** Organisations making substantial use of contact details within their business processes. These include DWP, HMRC, UKPS, DVLA, DfES, IND, GRO, GRO(S), eGU, ONS and Local Authorities via ODPM. See below for details of visits and interactions.
  - **Tier 2:** Other organisations with limited use of contact details. The CIP team have visited these organisations and described the scope of CIP, but have not identified any benefits worth pursuing further.
  - **Tier 3:** Remaining organisations. The CIP team have provided these organisations with a presentation of the scope of CIP and requested that they contact the team if they wished to explore the potential of CIP further.

### 3.2 Tier 1 stakeholder interactions

- 3.2.1 CIP policy and technical teams have in general carried out four rounds of meetings:
- **Initial meeting, May – August:** During which the scope of CIP was outlined and information gathered and requested regarding existing systems and current initiatives, initial discussion of business processes and events using contact details and potential benefits.
  - **Stage 1 meetings, August – November:** More detailed discussion of business processes, events and benefits arising from stage 1 technical options (models A-C). Also requested completion of data quality questionnaire.
  - **Stage 2 meetings, December – March:** Assessment of ID Cards as an adult population register and the additional functionality proposed by CIP. Completion of events / benefits, data questionnaire and current data sharing details.

- **Stage 3 meetings, May - June 2005:** Presentation and discussion of results from data trial and review of stakeholder profiles and data sharing information.

3.2.2 Each of the Tier 1 stakeholder profiles in the following sections summarises the following:

- Stakeholder business processes and events related to contact details
- Overview of stakeholder systems
- Overview of current initiatives
- Assessment of the data quality characteristics of existing data either through the data quality questionnaire or analysis of sample data within the data trial
- Existing data sharing with other organisations.

## 4. Driver and Vehicle Licensing Agency

### 4.1 Business processes and events

4.1.1 Within the Department for Transport (DfT), the Drivers, Vehicles and Operators Group (DVO) seeks to provide a one-stop shop for vehicle and driving information via the following agencies:

- Driver and Vehicle and Licensing Authority (DVLA)
- Vehicle Operator Services Agency (VOSA)
- Vehicle Certification Agency (VCA)
- Driving Standards Agency (DSA).

4.1.2 The Driver and Vehicle Licensing Agency (DVLA) was established as an Executive Agency in April 1990 and became a Trading Fund on 1 April 2004. DVLA manage vehicle registration and licensing and driver licensing within the UK, except for N Ireland, where a separate Agency, the Driver and Vehicle Licensing Northern Ireland, act on their behalf. The systems within both agencies are being tightly integrated. Current driver and vehicle systems are separate but are to be combined. DVLA are also looking at web enabling systems to allow citizen access. Roll out is planned for 2005 with full implementation by 2008.

4.1.3 The DVLA holds records of all holders of UK driving licences and all vehicle keepers (individuals and organisations). There are:

- 62 million driving licence records of which 40 million are active. Drivers must be 16+ years of age and may be of any nationality. Photo card driving licences must be renewed every 10 years (every 3 years for drivers over 70) and older, paper based, licences were issued for 25 years but are being replaced.
- 28 million licensed vehicles as at 2000. A registered vehicle keeper does not have to hold a driving licence and organisations register at least 40% of all vehicles. A vehicle licence may be for 6 or 12 months.

4.1.4 DVLA business processes and events are directed towards issuing driver and vehicle licenses. The main events during these processes are:

- Issue/renew driving licence
- Re-instate licence
- Register vehicle
- Vehicle VED payment
- Change address
- Vehicle registration enquiry (by police)
- Compliance/enforcement
- Vehicle recalls (a service to manufacturers).

- 4.1.5 DVLA consider their requirement to be contact data that is of high currency and verification. A high level of identity verification is desired when registering the keeper of a new vehicle. This is mainly done via Automated First Registration and Licensing (AFRL) system.

## 4.2 Current systems

- 4.2.1 Currently DVLA have separate vehicle and driver systems. The Driver system contains details of driver identity (including driver number), address, qualifications and entitlement, convictions and their related histories.
- 4.2.2 The vehicle system contains details of all registered vehicles, including those not currently in use (Statutory Off Road Notification (SORN)) and, optionally the vehicle mileage at renewal; together with the vehicle keeper's details – identity, address at which the vehicle is kept and, optionally, the vehicle keeper's driver number.
- 4.2.3 The Automated First Registration and Licensing (AFRL) is the system used by car retailers (holding a "Trade Plate") to register and licence new cars. These retailers have the responsibility for establishing the purchaser's identity.

## 4.3 Initiatives

- 4.3.1 DVLA have significant changes in progress, primarily driven by the DVO "one stop" customer service initiative, as follows:
- Customer Persistent Data Store (PDS) – common customer details for use across DVLA, VOSA and DSA. This will involve significant address matching of citizen data from existing systems.
  - Drivers Re-engineering Project (DRP):
    - Integrated vehicle and driver data using PDS
    - Electronic interaction with customers - 75% web transactions by 2005
    - 24 x 7 system operation
    - Improved identity and background checking (being developed in conjunction with UKPS)
    - Smart card driving licences.

## 4.4 Current data quality: Questionnaire responses

### ***DVLA driver database***

- 4.4.1 There are two sets of results for DVLA both have been reported based largely on detailed analysis of the datasets.
- 4.4.2 The drivers database holds details of the driver population aged 16 and over including non-license holders, this has 62 million records of which 40 million are live records. The degree of duplication within the dataset is not known although it

is accepted that there may be a small percentage of drivers with more than one record.

- 4.4.3 Address currency: CIP estimate this as a 62% probability of an address being current at any time. This could be improved by up to 10% when the data is combined with the vehicles database. See Appendix 3.3 for how this was derived.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100		100	100	100	100
<b>Address</b>	100	100	100	100	100	0
<b>Gender</b>	100			100		
<b>Date of Birth</b>	100		100	100	100	100
<b>Place of Birth</b>	Optional					Not known
<b>Date of Death</b>	Not known		Not known	100		

**DVLA driver database**

#### ***DVLA vehicle database***

- 4.4.4 The vehicles database holds details of keepers of vehicles, this has 30 million live records. Since the database is based around the vehicle it is entirely legitimate for one person to be listed more than once. Less than 60% of keepers are individuals as many vehicles are registered by companies.

- 4.4.5 Address currency: CIP estimate this as a minimum 90% probability of an address being current at any time. See Appendix 3.3 for how this was derived.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100		100	100	100	<100
<b>Address</b>	100	100	100	100	100	0
<b>Gender</b>	0			0		
<b>Date of Birth</b>	Optional		Not known	As for coverage	0	0
<b>Place of Birth</b>	Not Required					Not known
<b>Date of Death</b>	Not Required		Not known	Not known		Not known

**DVLA vehicle database**

## **4.5 Current data quality: Data trial results**

- 4.5.1 The DVLA driver data represented the second largest dataset analysed under the CIP trial, with 39,004 records in the sample.

### **Coverage**

- 4.5.2 Overall coverage for DVLA is equivalent to about 85% of the over 16 population (derived from the Electoral Roll and excluding ambiguous demographics s5 and s7<sup>1</sup>) DVLA to advise if data sample includes drivers who are dead, have expired licences, or no longer living in the UK. This is based on the average coverage of all demographics.

### **Field analysis**

- 4.5.3 The data comprised the following key fields:
- Unique reference number (Driver Number)
  - Date of Birth
  - Gender
  - Place of Birth
  - Surname and Forename fields
  - Alternative Surname and Forename fields
  - Separated Address fields
  - Account Creation Date
  - Last Update Date (initial results not available for Lot 1)
- 4.5.4 No information was available within supplied data on date of death or expired records.
- 4.5.5 Of the key fields listed above, nearly 100% of records are populated with the following exceptions:
- Alternative Surname and Forename fields, 16%
  - Place of Birth, 77%
  - Account Creation Date, 81%
- 4.5.6 Analysis of date of birth information shows good coverage of adults in the age range 16 to 85. There does appear to be a significant use of 1<sup>st</sup> January as a default birth date. In the full dataset, we would expect 47% +/- 13% of all 1<sup>st</sup> January dates to be defaults.

### **Identity duplication**

- 4.5.7 Comparing name, address and date of birth, a number of people were found who appear to have two separate driver records (driver numbers). 65 such records were found. Extrapolating to the full dataset, we would expect 0.17% +/- 0.04% of records to be duplicates. In a database of 40 million drivers, this would equate to between 52,000 and 84,000 duplicates. One driver with three records was also found.

---

<sup>1</sup> An outline of the demographic sets can be found in 3.2.16

### ***Demographic analysis***

- 4.5.8 Coverage for Scotland and Birmingham demographics is surprisingly low at only 10% and will be investigated further with the DVLA.
- 4.5.9 Demographic differences derived by individual column analysis were not particularly expected. However, demographic differences are apparent for gender and place of birth.
- 4.5.10 For gender overall, the data contains 57% males, which extrapolates to the full dataset as 57% +/- 0.6%. However, significant variation to this was seen for s6, s7 and s9. s6 (Wales) shows a male: female split closer to 50% than any other DVLA demographic, but still slightly dominated by males. S7 (Birmingham) and s9 (1<sup>st</sup> January) show a much larger percentage of males (c. 65%).
- 4.5.11 For place of birth, overall about 23% of records are null, and significant and large variance from this occurs for s5 (Scotland) (51% null), s7 (Birmingham) (62% null), s8 (DoB) (8% null) and s9 (DoB 1<sup>st</sup> Jan) (11% null).

### ***Address***

- 4.5.12 94.1% of DVLA address data is PAF compliant, which is above the 90% matching level at which the Post Office will start offering mailing discounts. While the total score was 95% compliant with PAF, only 6% were actually matched as “Verified Correct” as the DVLA generally omits the town name from its address format. The absence of this field in many DVLA records resulted in QAS making an automatic adjustment to the address format which, in turn, resulted in substantial number of records being classified as only a “Good Match” where such automatic town name insertions had been made.

## **4.6 Current data sharing**

- 4.6.1 Accuracy, driver checking and fitness to drive is facilitated by data feeds from:
- UKPS for correlation with passport details
  - ONS on births, marriages, deaths
  - Private sector credit reference details
  - Electoral register.
- 4.6.2 DVLA provide data to a variety of other organisations for enforcement purposes and checking on entitlements, such as VED exemption, driver licence check, offences, congestion charging and insurance checks. Some of these organisations are international, such as the Schengen information system and EUCARIS.

## 5. Department for Education and Skills

### 5.1 Business processes and events

5.1.1 The Department for Education and Skills is responsible for education and lifelong learning in England. It also has wider responsibilities for a range of policies, some of which it shares with other government departments, such as the Sure Start programme (shared with the Department for Work and Pensions), to ensure children and young people are safe, well and ready to learn.

5.1.2 While there is a great deal of use made of contact data by schools and other educational bodies, composite systems using such information are a much smaller set<sup>2</sup>. This means there is no real standardisation of contact information, even between schools<sup>3</sup>. From age 16 there is high fluidity of people moving across systems which will be extended to age 14 with the diversity in qualifications, such as GNVQs, which is occurring. There is currently a considerable need for appropriate data sharing in DfES. There appear to be three major projects that will use contact data across England and Wales in future. These are:

- Information Sharing and Assessment (ISA);
- Managing Information Across Projects (MIAP);
- Student Loans.

5.1.3 Of the above projects the most critical to DfES is probably ISA which was one of the recommendations from the report emanating from the Climbié Inquiry which was given legal ascent in the Children's Bill 2004. ISA will provide all the desired data for the portion of the population not currently covered by the National Identity Register (ISA will cover under 18's).

5.1.4 In the main the important data characteristic which is sought is coverage, followed by currency.

### 5.2 Current systems

5.2.1 The present situation is being progressed and in the absence of CIP, there are a number of local initiatives. DfES do not want to wait for the NIRNO and the fragmented models which will result in the meantime. DfES have ten applications in train to use the NINO.

### 5.3 Initiatives

5.3.1 DfES need to obtain better data to inform policy and intervention decisions. For example, at present there is a data delay difficulty in the Further Education area

---

<sup>2</sup> One such cluster is exam boards numbers, student loan a/c numbers and pupil/unique learner numbers.

<sup>3</sup> Although schools tend to use one of three software packages.

where information as to where age 16 students went after leaving school is not available until 15 months after they left.

- 5.3.2 DfES have stated their intent to have an on-line learners space where students would give others permission to share personal information. This information would include verified qualifications, personal learning logs (such as Record of Achievement) and a CV.
- 5.3.3 At the centre of the MIAP project is a learning data platform which has a new Unique Learner Number as the key identifier. Learners, education providers, national and local agencies will all feed into and utilise this platform and it will be possible for authorised people to view the assessment/achievement of students. The outcome of MIAP will therefore be a data sharing framework, used by a national register of providers working to common data definitions to provide a learner information service.
- 5.3.4 The Information Sharing and Assessment project is supported by legislation through the Children's Bill (2004) which enables the Secretary of State to require, through secondary legislation, that local authorities in England and Wales establish databases containing basic information about all children. The secondary legislation would be explicitly agreed by Parliament before it could take effect. Databases might be set up at a local, regional or national level. The purpose of the databases would be to facilitate the sharing of information between providers of children's services about the children they are working with, in order to safeguard their welfare and promote their well-being.

## 5.4 Current data quality

- 5.4.1 Student Loans 'Protocol' database: This database underlies the IT system used to administer higher education student financial support applications. It is managed by the Student Loans Company (SLC) on behalf of the DfES. Protocol is used by 172 Local Education Authorities in England and Wales to assess and approve student finance applications. SLC has a separate database called CLASS that handles payments and recoveries for all UK administrations which is outside the scope of this questionnaire.
- 5.4.2 The target population is eligible students applying for higher education to LEAs in England & Wales. Coverage is 95% at initial registration increasing to near 100% by graduation / drop-out and the NINO is used as the unique ID along with 'Automated Response Technology Number' (ART ID number)
- 5.4.3 . Students applying only for grants do not need to supply a NINO at all because they have no loan to repay, hence, the coverage will never be 100%.
- 5.4.4 There are approximately 5 million customer records in the Protocol database. However, these are not all "active" learners. Many are historic learners who have been pre-populated in Protocol from CLASS in case they return to HE. Many others are "sponsors" of learners e.g. parents and spouses who SLC communicate with even though they are not student finance applicants

themselves. The number of active learners in the database is approximately 800,000 in any one academic year.

5.4.5 1-2% of citizens may be entered more than once (possibly using different unique identifiers).

5.4.6 The data is kept up to date in Protocol whilst the person is applying for and receiving student finance support. Once they drop out or graduate there is no reason to keep the data up to date. Note, however, that the data in CLASS is kept up to date until the customer’s loan is fully repaid or written off.

5.4.7 Potential new and enhanced operational data sharing may improve data quality.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100		>99	100	>99	100 initially
<b>Address</b>	100	>99	>99	33	100	100 initially
<b>Gender</b>	100			100		
<b>Date of Birth</b>	100		100	100	>99	100
<b>Place of Birth</b>	Not Required					
<b>Date of Death</b>	Not Required					

**DfES: Student Loan Company**

## 5.5 Current data sharing

5.5.1 DfES is looking for better systems for exchanging data about students. There is an increasingly strong link between UCAS and Student Loans, then linking to the relevant institution. This is tied to the introduction of differential fees. Student Loans is also interested in exchanging data with the HMRC for the purpose of verifying parental income. At present this is a costly part of the operation.

## 6. Department for Work and Pensions

### 6.1 Business processes and events

6.1.1 The Department for Work and Pensions (DWP) provides services to four main customer groups identified through their National Insurance Number (NINO) or Child Reference Number (in the case of under 16 year olds):

- People of working age seeking work or unable to work due to incapacity
- Pensioners
- Families and children
- Disabled people and carers

6.1.2 The eligibility rules are complex.

6.1.3 To ensure correct entitlement, DWP holds a database of records for all UK residents including foreign nationals working, claiming or receiving benefits in the UK. Records are also kept for those living abroad who are in receipt of UK benefits or who are liable for UK National Insurance Contributions. The records of deceased persons are maintained on the database as they may be needed to claim benefits by other family members. Coverage also includes all children where child benefit has been claimed and many corporate organisations. DWP's coverage only extends to those people that the DWP need to deal with or know about. Not all the population is included, for example people in the 'black economy' or those who have never claimed benefit or worked in the UK will not be known to DWP. Similarly those who have never registered for a National Insurance Number with DWP will not have a database record.

6.1.4 There are a variety of business processes within DWP that use contact information. These processes include:

- Advising citizens of the range of benefits
- Understanding individual's circumstances to establish entitlement to benefits (which includes establishing identity as well as contact information)
- Calculating an assessment and making an award
- Actioning change of circumstances
- Finding employment for job seekers
- Calculating child support
- NINO allocation/refusal
- Tracking temporary addresses
- Dealing with disputes, investigating fraud.

6.1.5 As there are 4 main business areas within the Department, administering over 20 different benefits, each with its own events, ascertaining common events at which contact information is used is difficult.

6.1.6 In the main the important data characteristic that is sought is coverage of the population. This is important when it comes to notifying people of entitlements such as pensions. However there are sections of DWP where the currency of address data is more important, such as benefits linked to residency and in the case of CSA, finding people who have a liability to maintain children.

## 6.2 Current Systems

6.2.1 To help with determining eligibility, while DWP has several computer systems, all DWP personal information is currently held in just two large databases:

- Personal Details Computer System (PDCS)
- Departmental Central Index (DCI)

6.2.2 There are currently 84.5 million records in DCI and a subset of 35 million records in PDCS. The DCI system records for all people registered in the UK consists of:

- 47 million live adult records in UK
- 1 million live social security benefit recipients living abroad
- 15 million deceased records, date of death verified
- 1.5 million deceased records, date of death not verified
- 5.5 million, abroad not in receipt of benefit
- 2 million, inactive but not categorised
- 12.5 million child records

6.2.3 DCI holds a limited set of personal details and is used to trace customers not known to DWP (ie those not already in receipt of DWP benefits). The PDCS system holds less records (a sub set of DCI) but stores more information on corporate organisations and people known to the DWP through receiving benefits/pensions. The PDCS system also maintains an address history.

6.2.4 The address data in these systems are reasonably up to date and verified when people are claiming benefit. However, contact details are less comprehensive in other instances and in some cases may well be in excess of five years old as there is no business need to keep them up to date and the citizen is only required to report changes if in receipt of benefit. There are system-to-system links between DCI and the National Insurance Recording System (NIRS 2), which belongs to Her Majesty's Revenue and Customs (HMRC). This enables DWP to capture updates to personal details (addresses etc) for citizens who have

dealings with HMRC for tax purposes and HMRC to receive updates from DWP for those in receipt of benefit.

- 6.2.5 In recent years the data issues associated with duplicate NINOs have been actively addressed and ongoing work is substantially improving the quality of this dataset.

## 6.3 Initiatives

- 6.3.1 Both PDCS and DCI are to be replaced by a new Customer Information System (CIS) from March 2005 onwards. CIS is designed to provide a person centric view of the data within DWP customer facing systems and will allow other Stakeholders access to this information through an external gateway. The CIS dataset will hold relationships between customer and spouse, children and parents / carers, where these are derived from the status of the recipient of Disability Living Allowance for children and eventually CSA links.

- 6.3.2 CIS will have easier access to historical names and addresses currently held within PDCS. The switch to this new system will take a number of years but will assist DWP efficiency by reducing staffing costs. DCI and PDCS will be decommissioned in March 2006 and March 2007 respectively.

## 6.4 Current data quality questionnaire

- 6.4.1 The DCI database holds details of all in UK who have had tax/benefit interest and has over 84 million records. In the past year approximately 2,500 duplicate accounts have been identified on DCI through normal business processing and these accounts have been rectified in conjunction with HMRC tax records.

- 6.4.2 Address currency: There are over 11 million changes of addresses processed each year. These are notified by individual benefit systems and broadcast by DCI to other systems including HMRCs NIRS2. Changes such as these are based upon 'latest is best' principle and DWP systems accept changes with a later notification / start date than that already held.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100		<50	99	100	75
<b>Address</b>	96	67	74	67*	74	10
<b>Gender</b>	100			100		
<b>Date of Birth</b>	<100		<100	<100	<100	50
<b>Place of Birth</b>	Not required					Not required
<b>Date of Death</b>	18		18	18	18	17

## 6.5 Current data sharing

6.5.1 Much of the data received by DWP is in relation to the benefits or entitlements which it provides, such as:

- ONS on deaths and help with the Longitudinal Study
- HMRC about information on individuals in work, NI contribution records to calculate pensions and help with the Longitudinal Study
- The Met Office to help determine cold weather payments
- Audit Commission, NHS and local authorities in respect of possible fraud
- DVLA to help the CSA
- UKPS to aid matching and verification of identity
- The Post Office in respect of post codes

6.5.2 DWP has a strong matching service, MIDAS, which additionally receives information on topics such as the Construction Industry Scheme and from HMRC (to provide savings data) to help detect fraud as well as data such as Dental information from the NHS to match against DWP benefit data.

6.5.3 DWP provides information to many government departments because of its control (with HMRC) over the NINO, for example UKPS. It also shares information with local authorities in respect of housing benefit and confirms general benefit entitlement when an application is made for legal aid.

## 7. e-Government Unit

### 7.1 Business processes and events

- 7.1.1 The e-Government Unit works with departments to deliver efficiency savings while improving the delivery of public services by joining up electronic government services around the needs of customers. It also provides sponsorship of Information Assurance. The e-Government Unit takes on the majority of the work previously undertaken by the Office of the e-Envoy.
- 7.1.2 The e-Government Unit is responsible for leading the work on DirectGov website. This was launched in March 2004 as a new electronic service designed around the needs of the user, making it much easier to find and access government information and services electronically. It brings together information from across many Whitehall departments in one place, making it easier for people to find what they want from government, rather than having to search across several departmental sites. As well as government departments, the site links through to relevant third parties which can offer additional trusted advice and support.
- 7.1.3 The Government Gateway allows users to undertake secure electronic transactions with government. Registering with the Government Gateway enables sign up for any of the UK Government's services that are available over the Internet. It enables people to communicate and make transactions with government from a single point of entry.
- 7.1.4 Use of the Gateway requires registration. Following registration, the user has to enrol for each service they require, generally having to provide personal identifying data recognised by the service required (e.g. NINO for the Inland Revenue). In the current Gateway version, citizen addresses are not held, but are passed through to the secure printers for dispatch of identity number and PIN to individuals on registration or enrolment.
- 7.1.5 Although information is not stored in the Government Gateway itself, the main the important data characteristic which is sought is coverage.

### 7.2 Current systems

- 7.2.1 The UK Government Gateway uses a web service based authentication / authorisation model incorporating WS-Security, WS-Trust, WS-Profile. It is designed to scale to 60 – 100 million user accounts. It is based on internet connectivity, so it does not require new point to point connections to add additional organisations / systems. The existing system is able to generate multiple messages to required recipients from single authenticated source message, although this is capability does not appear to be used for any existing application.
- 7.2.2 The Gateway also incorporates:

- Payments engine
- Forms Engine
- Rules Engine
- Forms Store
- Circumstances engine

7.2.3 The last four of these enable the Gateway to capture data from users by means of forms. These forms can:

- Have fields pre-filled using data already held for the citizen using the Circumstances engine
- Have data structured to match target system requirements using the Rules engine.

7.2.4 Work is ongoing with respect to management of rules for different organisations (e.g. how these should be maintained, should EGU be providing standards or just enabling different departments to use their own).

### 7.3 Initiatives

7.3.1 The DirectGov website is due for an upgrade in approximately one year's time. After this release, citizens will be directed to use DirectGov for all services, while businesses will be directed towards Businesslink.

7.3.2 eGU is actively trying expand both range of services available and Government organisations (central and local) involved.

### 7.4 Current data sharing

7.4.1 As well as government departments, the DirectGov website links through to relevant third parties which can offer additional trusted advice and support

7.4.2 The Government Gateway uses internet based authentication<sup>4</sup> together with assisted form filling for onward processing and connection to departmental systems (e.g. enrolment to access service, submission of Inland Revenue tax returns).

7.4.3 Note that not all these organisations are citizen centric – some are business based (e.g. DEFRA: Procedure for the Electronic Application for Certificates from the HMI (PEACH) - for importers, exporters and processors of whole fresh produce into and out of the EU).

---

<sup>4</sup> XML messaging between Gateway and other systems

## 8. General Register Office (England, Wales and Scotland)

- 8.1.1 The GRO is part of the ONS and is responsible for ensuring the registration of all births, marriages and deaths in England and Wales, and for maintaining a central archive dating back to 1837.
- 8.1.2 There are equivalent offices for Scotland and Northern Ireland; GRO Scotland (GRO(S)) and GRO Northern Ireland respectively.

### 8.2 Business processes and events

- 8.2.1 The primary functions of the civil registration service are to register life-events (births, deaths and marriages, plus civil partnerships from late 2005), to maintain archives of the records of those events and to provide facilities for civil marriage and partnership. The service is delivered via a tripartite partnership of GRO, local authorities and registration officers.
- 8.2.2 Key events are:
- Record birth: At the point of birth a NHS number is allocated to each child, this is then used to identify the child when registering the child's name with the local Registrar.
  - Record death: A doctor's certificate is presented to the Registrar usually by the next of kin and a formal death certificate issued.

### 8.3 Current systems

- 8.3.1 The GRO holds records on key life events that take place in England and Wales, including births, marriages and deaths. These registers date back to 1837. Electronic records start from 1993-4. Once a registry entry is recorded, contact details are not updated under normal circumstances.

### 8.4 Initiatives

- 8.4.1 "Modernising Civil Registration" is a major initiative being implemented by the Civil Registration Review Programme (CRRP).
- 8.4.2 The Civil Registration Review (CRR) began in 1998. Its principal aim is to reform the civil registration service in England and Wales.
- 8.4.3 The CRRP is essentially about electronic data sharing of registration events with other government organisations. Data sharing facilitated by CRRP is listed in Annex 2E: Current data sharing across government.
- 8.4.4 The Government published a White Paper "Civil Registration: Vital Change" in January 2002 setting out its proposals for the reform of the service and

announced that reform will be made using the order-making powers of the Regulatory Reform Act 2001. Proposals in the form of a draft Regulatory Reform Order, that would have reformed the law on registering births and deaths and the structure of the local registration service, were presented to Parliament in July 2004.

8.4.5 The Commons' Regulatory Reform Committee and the Lords' Delegated Powers and Regulatory Reform Committee decided that the draft Order could not proceed to the second stage of scrutiny as it was an inappropriate use of the powers. The Government remains committed to the modernisation of the registration service and is actively looking for ways of delivering the key elements of reform. Once the path to reform has become clearer, it will be possible to make more specific recommendations about how the NIR should interact with CRR.

8.4.6 ONS are currently working on the delivery of a new database, RON – Registration On – Line and digitisation of their existing stock of registration records. This ultimately will provide a modern database, containing a comprehensive set of birth, death and marriage data for all events registered in England and Wales.

## 8.5 Current data quality: Data trial results

### GRO birth data

8.5.1 Refer to detailed analysis from data trial in Appendix B. Coverage: Only contains data since 1993.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100	99	95	99	99	N/A
<b>Address</b>	100		74	67*	57	0
<b>Gender</b>	100		100	100	100	
<b>Date of Birth</b>	100	100	<100	<100	100	N/A
<b>Place of Birth</b>	100					N/A

### GRO(Scotland) birth data

8.5.2 Refer to detailed analysis from data trial in Appendix C. Coverage: Only contains data since 1974.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100	99	92	100	100	N/A
<b>Address</b>	100		88	67*	71	0
<b>Gender</b>	100		100	100	100	
<b>Date of Birth</b>	100	97	<100	<100	100	N/A
<b>Place of Birth</b>	18					N/A

## 8.6 Current data sharing

8.6.1 ONS is liaising with public sector stakeholders to determine data requirements for receiving birth, death and marriage information, the position is summarised in the table below:

Public Sector Stakeholder	Current Data Provided	Plans for additional data requirements
DVLA		No current plans – current legislation does not enable ONS to provide data automatically to DVLA
DWP	Deaths	ONS currently discussing DWP requirements for access to Marriage and Civil Partnership data.
NHS	Births and Deaths	No plans to share Marriage data.
HMRC		Discussions are underway to explore how child births and deaths can be reported to HMRC. Existing legislation should cover this requirement.
DfES	Registrars notify Local Education Authorities of births	Discussions are underway with DfES on the possibility of providing a central birth notification. The legislative position is still to be determined.
Home Office – ID Cards Scheme/UKPS		Secondary legislation to the ID Cards Bill may enable the Registrar General to provide registration data to UKPS and the Home Office

## 9. Immigration and Nationality Directorate

### 9.1 Business processes and events

- 9.1.1 The Immigration and Nationality Directorate is part of the Home Office. IND are responsible for immigration control for the UK. They also administer permissions to stay, citizenship and asylum. For those asylum seekers that cannot support themselves, National Asylum Support Service (NASS), administers the accommodating of Asylum seekers, either allocating them to Local Authorities or detention centres for temporary housing whilst their cases are investigated. An agency of IND is UKPS.
- 9.1.2 Key business processes involving permissions to stay use contact details. For the majority of people these details are good and updated when people come to renew these permissions. Difficulties emerge when IND wants to follow up persons with leave to stay in UK for fixed period or failed Asylum Seekers. People “disappear” and are difficult for the enforcement teams to find. Very often they also want to identify others living at same address as target subject.
- 9.1.3 The important data characteristic which is sought is currency (of address).

### 9.2 Current systems

- 9.2.1 A recent competition has been won by Atos Origin which provides for the replacement of all ICT infrastructure services provided to IND staff, plus a controlled range of enhancements to address significant deficiencies identified for action in the IND IT Strategy, particularly in respect of IT service integration. IND have contracted with the new supplier for application systems development services but has reserved the right to select a different supplier as it deems appropriate. The previous contract with Siemens for the management, maintenance and development of the Team Based Caseworking (TBC) network and ICT infrastructure services expired in October 2003.

### 9.3 Initiatives

- 9.3.1 e-Borders (electronic borders) is a cross-cutting initiative co-ordinated by the Home Office in partnership with key border control, law enforcement and intelligence agencies. The e-Borders system will identify people who have boarded transport destined for the UK, check them automatically against databases of individuals who pose a security risk, and keep a simple electronic record of entry into the country. The system will also enable authorities to record people leaving the UK, helping to identify those who overstay.
- 9.3.2 Project Semaphore is the first stage in the Government's e-Borders programme. It is a £15 million pilot scheme which initially will target six million passengers a year travelling on a number of international air routes to and from the UK. It will provide a comprehensive passenger movement audit trail. Semaphore will test

and confirm the technical and business process design for the main e-Borders programme as well deliver immediate operational improvements across participating agencies. It will enable tighter controls of those who may pose a security risk and ensure we can have clearer records of those entering and leaving the UK.

- 9.3.3 Project IRIS (Iris Recognition Immigration System) is a secure, automated border entry system using iris recognition technology to speed up the admission of pre-assessed bona-fide travellers is being piloted and should be fully operational by the summer of 2005

## **9.4 Current data sharing**

- 9.4.1 Address change details on Asylum Seekers are exchanged between Local Authorities and NASS by CD.
- 9.4.2 The Warnings Index System is checked when a person enters the country (refused entry is notified to HM Customs, Intelligence Services and others). UKPS access this file as part of their pre issue or renewal process for passports.

## 10. HM Revenue and Customs (formerly Inland Revenue)

### 10.1 Business processes and events

10.1.1 The HM Revenue and Customs (HMRC) is responsible for the administration of:

- Income tax
- Corporation tax
- Capital Gains tax
- Petroleum revenue tax
- Inheritance tax
- National Insurance contributions
- Stamp Duties
- Charities
- Student loan collections

10.1.2 The HMRC is also responsible for the payment of:

- Child tax credits
- Child Benefit
- Child Trust Fund

10.1.3 Citizens are identified through the use of the NINO, Temporary NINO or the Unique Tax Reference number. In the majority of cases, the HM Revenue and Customs is also responsible for collecting student loans. It also provides a range of electronic services to people and businesses and plays a role in enabling charity through payroll giving.

10.1.4 The list of main HMRC events concerning citizen contact information are:

- Employer notification of individual starting/ceasing employment
- Advise citizen of tax code (individual mailing)
- Send citizen tax return (individual mailing)
- Investigate non-compliance (tax returns)
- Issuing reminders
- Notify tax liability/assessment/interest and penalties
- Make or demand payment
- Change of circumstances
- Disputes (incl. appeals)
- Exception processing and tracing
- Verification of the e-channel (applies to all) - applying for the government gateway
- Clerical cases.

- 10.1.5 HMRC consider their key requirements in contact details are for data with characteristics of good currency and verification.

## 10.2 Current systems

- 10.2.1 The key HMRC systems are experiencing capacity issues which are being addressed through a planned programme of work. For example, the Computer Operation of PAYE system (COP) was built on a series of regional databases, but is being replaced by the Modernisation of PAYE Processes for Customers project (MPPC). Again the Child Benefit system holds data on individuals receiving child benefit and is being replaced with CB2 as a result of the new Child Trust Fund (CTF). The Child Benefit system interacts with the Department for Work and Pension's (DWP) Departmental Central Index (DCI)<sup>5</sup> from which it draws child reference numbers (CRNs) that become an individual's NINO from age 16 onwards.
- 10.2.2 The National Insurance Recording System 2 (NIRS2) is designed to collect contributions; hold more than 65 million individual contribution records; calculate contributory benefits; provide data to other government agencies; and pay age related rebates to Occupational and Personal Pension schemes. NIRS2 is one of the biggest IT systems in Europe, with over fourteen million lines of code.

## 10.3 Initiatives

- 10.3.1 HMRC is establishing a Citizen ID Framework (CID) and an Address Framework to standardise access to name/address information across the majority of systems<sup>6</sup> in HMRC. This work will be completed by April 2006. The Address framework provides a means of aiding postal delivery. This provides a consistent use of addresses across HMRC and helps to maximise postal discounts. It provides a means to match incoming address data against an authoritative source, such as Royal Mail's Postcode Address File (PAF).

## 10.4 Current data quality: Questionnaire response

- 10.4.1 These results relate to the Citizen Identification / Address Frameworks database which holds 60 million records. This covers adults with liability to UK tax or entitlement to claim tax credits or child benefit and children for whom child tax credit or child benefit is claimed. It is believed that within this number there are about 4 million with duplicate temporary or permanent NINOs. HMRC reported that these results are based on feedback about the dataset.

---

<sup>5</sup> DCI is itself being replaced by the new Customer Information System (CIS)

<sup>6</sup> The HMRC has around 300 existing IT systems

10.4.2 Address currency: HMRC reported that currency of address data is largely dependent on customer notification of a change, which happens either at time of change or later. Address is taken as current and used when employer provides it as part of start of employment information. Employer obtains information from employee and provides it to HMRC at the time that employment starts.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100		61	48	90	9
<b>Address</b>	99	99	94	98	99	0
<b>Gender</b>	90			81		
<b>Date of Birth</b>	85		85	85	0	17
<b>Place of Birth</b>	Not required					Not required
<b>Date of Death</b>	7		7	7	0	0

## 10.5 Current data quality: Data trial results

10.5.1 The HMRC data represented the largest dataset analysed under the CIP trial, with 93,583 different records in the sample.

### **Coverage**

10.5.2 Overall coverage for HMRC is equivalent to about 83% of the population, based on the average of the estimated coverage for all nine demographics and will include foreign nationals who have worked in the UK and now left.

### **Field analysis**

10.5.3 The data comprised the following key fields:

- Unique reference number (NINO)
- Date of Birth
- Date of Death
- Gender
- Date of Entry (around the date on which person was 16)
- Date of Registration (date person registered on NIRS)
- Name with separate Surname and Forename fields
- Name Last Update Date
- Address with separated Address fields
- Address Last Update Date

10.5.4 Analysis of Date of Birth information shows good coverage of adults from age 16. There appears to be a significant use of 1<sup>st</sup> January as a default birth date. In the full dataset, we would expect 54% +/- 8% of all 1<sup>st</sup> January dates to be defaults.

### ***Identity duplication***

10.5.5 Comparing name, address and date of birth, a number of people were found who appear to have two separate records (National Insurance Numbers). 66 such records were found. Extrapolating to the full dataset, we would expect 0.071% +/- 0.017% of records to be duplicates. In a database of 40 million people, this would equate to between 21,600 and 35,200 people with duplicate (i.e. two) records (National Insurance Numbers).

### ***Demographic analysis***

10.5.6 Demographic differences derived by Individual Column Analysis were not particularly expected. However, demographic differences are apparent for Gender.

10.5.7 For Gender overall, the data contains 51.3% males, which extrapolates to the full dataset as 51.3% +/- 0.4%. However, significant variation to this was seen for or s1 (name), s2 (name), s7 (Birmingham) and s9 (DoB 1/1). For s1 and s2, this is reversed, with female dominance of 52% and 54% respectively. S7 and s9, somewhat in support of DVLA, shows an increased male dominance, though only marginal at 52.4% and 59.4% respectively.

### ***Address***

10.5.8 89% of addresses complied with PAF but this was the only dataset to include all historical addresses and the overall score for this dataset suffered from the obsolescent nature of some of its addresses.

## **10.6 Current data sharing**

10.6.1 HMRC receives data from a number of departments and local authorities:

- DVLA on enforcement - medical fees summary and taxes management;
- DWP in respect of National Insurance Numbers (NINOs), child reference numbers and changes to the National Insurance system;
- ONS (GRO) for notification of births in connection with child benefit claims;
- Local Authorities for Child Benefit and Tax Credit purposes;
- Customs – integration to EU (corporate);

10.6.2 HMRC provide data to several sources to check on eligibility for Child Benefit, amount of student loans, tax credits and to help with statistical and research analysis.

## 11. National Health Service

### 11.1 Business processes and events

11.1.1 The NHS consists of a range of health and social care services delivered by a fairly diverse set of organisations. Each of the four UK countries has its own, separately administered NHS, although there is close working between the four administrations to ensure, as far as possible, seamless delivery of services to patients and citizens. The Department of Health sets overall policy on health issues for the NHS in England, and provides accountability to Parliament for monies spent by the NHS, and for the overall delivery of health service policy. It is the responsibility of local Primary Care Trusts (PCTs), which are autonomous, to provide health services to the general public.

11.1.2 The National Health Service (NHS) was set up in 1948 to provide healthcare for all citizens (denizens) based on need and not the ability to pay; it is funded by the taxpayer. It is recognised as one of the best health services in the world by the World Health Organisation.

11.1.3 The NHS is engaged on a major modernisation programme to reflect the changing demands and aspirations of citizens in the 21<sup>st</sup> century. The key drivers for this programme of change are the “NHS Plan”, which reaffirmed the central ethos of the NHS as being to provide health and social care services that are responsive to the needs and wishes of patients and their carers; and the “Wanless” report which identified the need for major investment in the NHS, particularly in terms of IT, if the NHS was to continue to deliver the high standards of service that citizens expected. A key policy goal is to ensure that the delivery of health and social care services are centred around patients, and offering them more choice over how, when and where they wish to be treated. The National Programme for IT for the NHS in England – one of the world’s largest civil IT programmes – is creating a comprehensive information infrastructure at the heart of which is the NHS Care Records Service (NHS CRS), which includes an index of all patients of the NHS in England.

11.1.4 The processes of the NHS are directly concerned with delivering health and social care. As a patient receiving advice or treatment particular events are:

- GP Registration
- Presenting at a GP, hospital or walk-in centre
- Receiving care at home or in the community
- [Interactions](#) with the private and voluntary sector
- [Interactions](#) with NHS Direct
- Screening.

11.1.5 As a patient receiving medicines events are:

- Prescription entitlement check
- Prescription verification.

11.1.6 The key desired data characteristics for patient details are generally uniqueness, although for screening programmes currency and validity are also important. Uniqueness is important because of the need to match patients with their own records.

## 11.2 Current systems

11.2.1 The NHS Strategic Tracing Service (NSTS) holds key administrative information including: NHS number, name, date of birth, sex, date of death (where applicable) and details of all GP-registered patients, GP details and practice addresses. This information is provided by GPs (via the local PCT) and the Registrar of Births and Deaths (via the ONS). It is comprehensive and covers all patients in England and Wales. The main function of the NSTS is to enable NHS organisations to trace the NHS numbers and confirm the associated personal details of their patients thus improving data quality. The NSTS will be phased out over the next 12-18 months to be replaced by the “Personal Demographics Service” (PDS) element of the NHS Care Records Service.

11.2.2 The NHS Central Register (NHSCR) (or CHRIS), run by the ONS for the DH, compiles and maintains a computerised central record of those patients who are registered with a NHS GP in England, Wales and the Isle of Man. The NHSCR contains over 60 million records from all PCTs.

11.2.3 The 88 National Health Applications & Infrastructure Services (NHAIS) (Exeter) systems are used by all PCTs in England and Wales to link to the NHSCR, and for the administration of cancer screening call/recall programmes and to deal with patient registration and contractor payments.

11.2.4 The NHS also has other non central systems:

- GP systems – individual systems which are either stand alone and feed into the NHAIS systems, or are NHS CRS enable and are able to connect to the PDS.
- Acute Trust PAS
- Maternity systems
- Child health
- CIS – NN4B

## 11.3 Initiatives

11.3.1 Connecting for Health, an agency of the Department of Health, is responsible for delivering the National Programme for IT (NPfIT) for the NHS in England. A core element of the NHS Care Record Service, which will lead to England’s NHS having an integrated, electronic record management service for the first time. The

NHS CRS is an electronic record management service that allows authorised care professionals secure access to an individual's NHS Care Record 24 hours a day, seven days a week, whether they work in GP practices, hospitals, community health or social services. The NHS Care Record is a means of ensuring that the details of a patient's care and treatment are held in a single, easily accessible, electronic record. One of the principal benefits of the NHS CRS is that the right information will be available to the right people at the right time – securely and confidentially.

11.3.2 The Personal Demographic Service (PDS) element of the NHS CRS will be the definitive source of patient demographics and will include data such as name, NHS number, sex, data of birth and address. PDS Phase 1 Release 1(P1R1) went live December 2003 and was initially a bulk data transfer from the NSTS, from which it continues to receive daily updates. PDS P1R2 is due to go live in 2005 and will support multiple addresses and issue NHS Numbers directly.

11.3.3 While there are strict protocols governing access to sensitive clinical information, the demographic information held on the PDS is regarded as less sensitive and does not have the same range of complex range of permission matrices, although access to PDS is still restricted to appropriately authorised NHS personnel. Certain types of records (e.g. adoption) are subject to special controls to comply with legal requirements.

## 11.4 Current data quality questionnaire

11.4.1 Awaiting reply.

## 11.5 Current data sharing

11.5.1 The NHS takes data from

- GRO births and deaths on a weekly basis
- DfES and DWP
- HMRC on child benefit matters
- DVLA in respect of organ donors
- MOD where the NHS handles transfer of medical records of dependants of service personnel between Service Medical Units and civilian GPs on behalf of the MoD.

11.5.2 The NHS provides only a limited amount of data to other organisations, because of legal constraints over patient confidentiality, but it does so in areas where there is a statutory or other legal requirement to do so, such as to the GRO in notifying births and deaths, to the Prison service, to the Audit Commission and DWP for purposes of investigating fraud and resource allocation between GPs. The NHS also has a duty to inform local authorities under the Public Health act of notifiable diseases and food poisoning.

## 12. Office for National Statistics

### 12.1 Business processes and events

12.1.1 The Office for National Statistics (ONS) is the government department that provides UK statistical and registration services. ONS is responsible for producing a wide range of key economic and social statistics which are used by policy makers across government to create evidence-based policies and monitor performance against them. The Office also builds and maintains data sources both for itself and for its business and research customers. It makes statistics available so that everyone can easily assess the state of the nation, the performance of government and their own position.

12.1.2 The Office also incorporates the General Register Office for England and Wales (GRO) see section 8. The GRO is responsible for ensuring the registration of all births, marriages and deaths in England and Wales, and for maintaining a central archive dating back to 1837.

12.1.3 The main process that uses citizen contact data within ONS is the population census, where there would be a cost saving from not running a full Census. Administrative data (from NIR/DWP/NHS) would provide population census statistics and smaller targeted surveys would fill in the gaps. This approach would also enable an ongoing population census and by using administrative data would allow a population census to be undertaken yearly rather than every ten years. This would give more current population census statistics.

12.1.4 In the main the important data characteristic which is sought is coverage

### 12.2 Current systems

12.2.1 Census 2001 data was collected via Census forms<sup>7</sup>. A similar approach is planned for 2011.

### 12.3 Initiatives

12.3.1 A population census can be drawn from extracting basic fields from administrative data sources. Possible primary data sources include:

- National Identity Register (ID Cards);
- DWP
- NHS

12.3.2 Additional sources with less coverage could be used to enhance the census results further, e.g. DVLA, UKPS, CORE (Electoral Register). Data matching and

---

<sup>7</sup> See (<http://www.statistics.gov.uk/census2001/censusform.asp>)

linking would need to be used where multiple data sources are used and do not share a common key.

## 12.4 Current data sharing

12.4.1 ONS obtain a wide variety of statistical information from a number of government sources.

12.4.2 The Integrated Population Statistical System (IPSS) is a feasibility research project looking to deliver a 'proof of concept' in 2007. The primary objectives are:

- To establish whether administrative data from different sources can be successfully linked on the premise of an IPSS to produce statistics, that will in some areas replace the need for a census;
- To establish statistical benefits of an IPSS database to ONS and broader Government Statistical Services (GSS);
- To assess the benefits of the linkage to DWP, Population and Demography Census 2011.

12.4.3 The study will perform a trial using the following data-sets:

- Census 2001;
- DWP's linked database;
- NHSCR data (currently provided for migration estimates)

## 13. UK Passport Service

### 13.1 Business processes and events

13.1.1 The UK Passport Service (UKPS) was established as an Executive Agency of the Home Office on 2 April 1991. The United Kingdom Passport Service (UKPS) aims to assist new applicants and existing passport holders by providing information and online facilities for all aspects of application, renewal and amendments of passports for British nationals resident in the UK. Its vision is “To focus on stronger identity authentication for the purpose of issuing passports and providing identity services”. The major challenge now faced by UKPS is to keep one step ahead of the difficulties posed by increasing fraud.

13.1.2 The UKPS typically holds information on all citizens in the UK and abroad (through the FCO) who have at some time in their lives held a UK passport. The UKPS database holds about 60M records in total (assumed to be expired and current passports), which includes citizens abroad and issues 4.5M passports per year. 80% of citizens with British nationality (including ex pats) hold a British passport, this includes 80% of children. However 25% of UK residents have never had a UK passport i.e. have never travelled or held a foreign passport. Based on this figure, approximately 35.8million adults and 8.93 million children within the country hold a passport [check figures].

13.1.3 The processes of the UKPS are directly concerned with the issue of passports:

- Receive passport application
- Establish identity and contact details from supplied documents and external checks – a biographic footprint
- Issue passport
- Reissue passport where necessary
- Investigate Fraud.

13.1.4 The key desired data characteristics for contact details are validity and high currency.

13.1.5 During the peak season in 2003, UKPS employed 2,890 full-time equivalent staff, over 90% of whom work in regional offices. The work of UKPS is very seasonal and the peak load is about 185,000 applications per week in high season and 30,000 per week in the low season. This produces considerable imbalance in the workload for staff, leading to seasonal employment and inefficiencies. This arises because addresses are not maintained on passport records and at the time of renewal can be ten years out of date, so renewal reminders are not sent.

## 13.2 Current systems

13.2.1 The current system (PASS) was introduced in 1998, but passports were still issued from the old system up to the end of 2001. The old system only recorded the book (passport number) and name(s) and relationships between children and the adult on whose passport they appeared. PASS includes digitised photographs and signatures and does not contain relationships as all children now have their own passport. PASS is limited to the issue of passports to UK residents. The Foreign Office is responsible for passport issues abroad and passes the information to PASS (0.5M per year).

13.2.2 PASS is passport rather than person centric i.e. it allows multiple passport numbers for one person (renewals, loss etc).

## 13.3 Initiatives

13.3.1 UKPS have a major programme of projects (the Integrated Change Programme) in place to improve their ability to prevent and detect fraud. There are eight key areas within this change programme. These are:

- Authentication by Interview (AbI)
- Personal Identification Project (PIP)
- e-Passport
- Births Marriages and Deaths Online (BMD)
- Naturalisation Online
- Facial Recognition
- Lost, Stolen and Recovered (LSR) passports; and
- Second Biometric.

13.3.2 These are supported by a number of supporting and enabling projects, for example Enterprise Data Warehouse (EDW).

## 13.4 Current data quality: Questionnaire response

13.4.1 UKPS (Main) database - awaiting completed response.

13.4.2 These results relate to the Main Index database which holds 70 million records. This covers all citizens who wish to travel abroad. There will be duplicate passport records within the database as the unique identifier is the passport number not the identity. Consequently the number of records will relate to the number of books an individual has been issued, with not the number of individuals. For each passport there are two associated record types (a passport record and an application record) each with a unique identifier. UKPS results are largely based on anecdotal evidence about the dataset.

## 13.5 Current data quality: Data trial results

13.5.1 Address currency: CIP estimate that there is a 60% probability of an address being current at any time. See Annex 2B Section 5 for the approach adopted in deriving this result.

	Coverage	Consistency	Completeness	Formatting	Validity	Verification
<b>Name</b>	100	99	83	90	91	100
<b>Address</b>	??	??	99	100	87	By process
<b>Gender</b>	100	99		100	100	
<b>Date of Birth</b>	100	100	100	100	100	100 ?
<b>Place of Birth</b>	100					-
<b>Date of Death</b>	0					

13.5.2 The UKPS data sample is from the PASS system and consists of 24,529 different records in the sample (excluding the 3 records that were sampled twice via the various demographics).

### **Coverage**

13.5.3 Overall coverage for UKPS (PASS) appears to be about 44% of the population. The data covers a period from 1998, i.e. PASS has been in operation for about 6 years, and given a 10-year renewal period for the passport, no more than 60% of the population would currently be expected within the database. The implication is that 73% of the population are passport holders.

- Unusually low coverage is seen for s5 Scotland, suggesting lower numbers of passport holders in this region than the national average.

### **Field analysis**

13.5.4 The data comprised the following key fields:

- Unique reference number (Passport Number)
- Date of Birth
- Gender
- Place of Birth
- Name with separate Surname and Forename fields
- Address with separated Address fields
- Account Creation Date
- Last Update Date

13.5.5 In addition, the following fields were included:

- Former Name

- Second Address fields

13.5.6 Analysis of Date of Birth information shows good coverage of adults from age 16, and an increasing coverage of young children. There does appear to be a significant use of 1<sup>st</sup> January as a default birth date. In the full dataset, we would expect 33% +/- 24% of all 1<sup>st</sup> January dates to be defaults.

### ***Identity duplication***

13.5.7 Comparing name, address and date of birth, a number of people were found who appear to have two or more passport records. This is to be expected for renewals, legitimate duplicates, etc. 406 such records were found (1.68%).

Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of:

More than one record:	1.68% +/- 2sd = 1.68% +/- 0.08%
Two records:	1.65% +/- 2sd = 1.65% +/- 0.08%
Three records:	0.03% +/- 2sd = 0.03% +/- 0.01%

13.5.8 In a database of 24 million people, this equates to:

More than one record	between 384,000 and 422,000 people
Two records	between 376,000 and 415,000 people
Three records	between 48,000 and 96,000 people

### ***Demographic analysis***

13.5.9 Demographic differences derived by Individual Column Analysis were not particularly expected. However, demographic differences are apparent for Gender.

13.5.10 For Gender overall, the data contains 48% males, which extrapolates to the full dataset as 48.0% +/- 0.3%. However, significant variation to this was seen for s9 (DoB 1/1), where the rate drops to only 41% +/- 1.4%. Note that this is in contrast to the other datasets which show a predominantly male ratio in the s9 demographic.

## **13.6 Current data sharing**

13.6.1 UKPS receives data from a number of sources which are designed to keep the passport system robust:

- FCO for passports issued abroad<sup>8</sup>
- DWP in relation to National Insurance numbers
- ONS about births and deaths notification and ELVIS<sup>9</sup>

---

<sup>8</sup> Using Genie a system similar to PASS. Emergency passports are currently paper and a photograph, although there are plans for a more secure document.

<sup>9</sup> ELVIS is the Events Linkage Verification and Information System which eliminates some types of fraud e.g. passport applications for citizens who have died before their 18th birthday.

- IND warnings index
- Courts in respect of child protection orders and banned football fans
- Police and Interpol on stop files and lost stolen and recovered (LSR) passports
- Under the PIP pilot, information is received with DVLA about a driver's record.

13.6.2 UKPS shares information with other government departments and the private sector to ensure that identity authentication services are as robust as possible. For example in respect of LSR, UKPS shares this information with IND, FCO, the Police and Interpol. An Office browser system allows access across the GSI to 95% of the data held by UKPS (both people and passports) including matching against stop files. This Office browser system (Omnibase) is used by DVLA, Criminal Records Bureau, Metropolitan Police, Special Branch, IND, Customs and Excise and DWP. Others (including from other countries) will join this service through Project Semaphore. Separately the DVLA have a pilot system working where they supply a passport number and data is returned across the GSI.

## Appendix A: Data assessment - DVLA

---

## 1. Data structure

1.1.1 DVLA data was delivered in nine files, one for each demographic. The file format was fixed length with commas separating each field. There were no quotes around the textual data, and several examples of embedded commas were found in the Address Line 1 column. This meant that the data could not be treated as comma delimited, but had to be treated as fixed length.

1.1.2 The data contained the following information:

- Unique reference number (Driver Number)
- Date of Birth
- Gender
- Place of Birth
- Surname and Forename fields
- Separated Address fields
- Account Creation Date
- Last Update Date (not available for Lot 1)

1.1.3 There was only the one address, but the data did include the following

- Alternative Name fields

1.1.4 Data extract issues were found with the data:

- The files containing the data for samples s1 and s2 were found to have omitted the column for AlternativeName.Forename. Replacements for these were supplied by DVLA.
- All files (s1-s9) had an additional unprintable character at the start of the AlternativeName.Surname column. This is the ASCII value 0x01 (ASCII Start of Heading character). It assumed that this is an artefact of the data extraction process and is not contained within the live DVLA database.
- There was no Last Update Date data. This column is of particular interest to the trial, in order to understand the currency of the DVLA data. DVLA ran a new extract including this information, however, this was not in time for the full Lot 1 analysis as published in version 1.0 of this document. The Last Update Date from this file has now been analysed as described under paragraph 44,

- Individual field analysis, Last Update Date. However, yet further extract issues were found with this file:
  - ASCII SOH problem now gone away
  - ASCII NUL characters used to pad short postcode fields instead of spaces, causing unexpected truncation of the record in our data processing tools (NUL is taken as an end of text character in the 'C' programming language). Special handling was required to remove the ASCII NUL prior to processing the data.
  - Garbage data in the "Last Update Date" column (c. 1300 records).

1.1.5 The nine DVLA files containing s1-s9 data were amalgamated in a single s0 data file containing 39013 records.

1.1.6 Postal town data was not sent because the DVLA derive the town from the postcode. Special dummy postcodes AA88 for post towns in Welsh and AA89 for postcodes with changed post towns are used uniquely by DVLA. DVLA insert 0 to postcodes of format AAN\*\*\* to show AA0N\*\*\*, insert a space to change formats of ANN\*\*\* to A NN\*\*\* and insert a 0 and a blank to changing format AN\*\*\* to A 0N\*\*\*. Address data was more consistent than for UKPS.

## 2. Statistical summary

2.1.1 Total records: 39013 (Initial analysis)

2.1.2 Total records: 39196 (Last Update Date analysis) (of which 81 are 2005, and 50 are 2004). The remaining additional records are distributed throughout the demographics, and potentially have been introduced by name or address changes bringing a net inflow of new records into the extract criteria.

Demographic		Frequency	% of s0
s1	Typical Dataset by name	3203	8.2
s2	Typical Dataset by name	2651	6.8
s3	Typical suburban dataset by geographic area (postcode and area name)	8648	22.2
	AA89	10	
	Incomplete Postcode (with *)	44	
s4	Covers name issues and address issues on houses that have been converted into flats. (postcode)	12885	33
	Incomplete Postcode (with *)	1	0
s5	Covers a rural area in Scotland (postcode)	355	0.9
s6	Covers issues around Welsh names and addresses (postcode and area name)	4080	10.4
	Incomplete Postcode (with *)	18	0
	AA88	1	0
s7	Covers issues related to high density urban areas	547	1.4

	and high rise flat blocks		
s8	Dataset by specific date of birth	2485	6.4
s9	Covers issues around nominated date of birth being 1 <sup>st</sup> January	4159	10.7

### 3. Individual field analysis

#### 3.1 Driver number

3.1.1 This not null unique identifier shows 9 people intersecting two demographics.

Intersected Demographic	Frequency
s1-s9	1
s2-s4	1
s2-s8	1
s3-s9	1
s4-s9	4
s6-s1	1

#### 3.2 Date of birth

3.2.1 In the format CCYYMMDD. Some records showed year or year and month only, with day or month and day values set to zero, presumably to indicate that the missing data is not known.

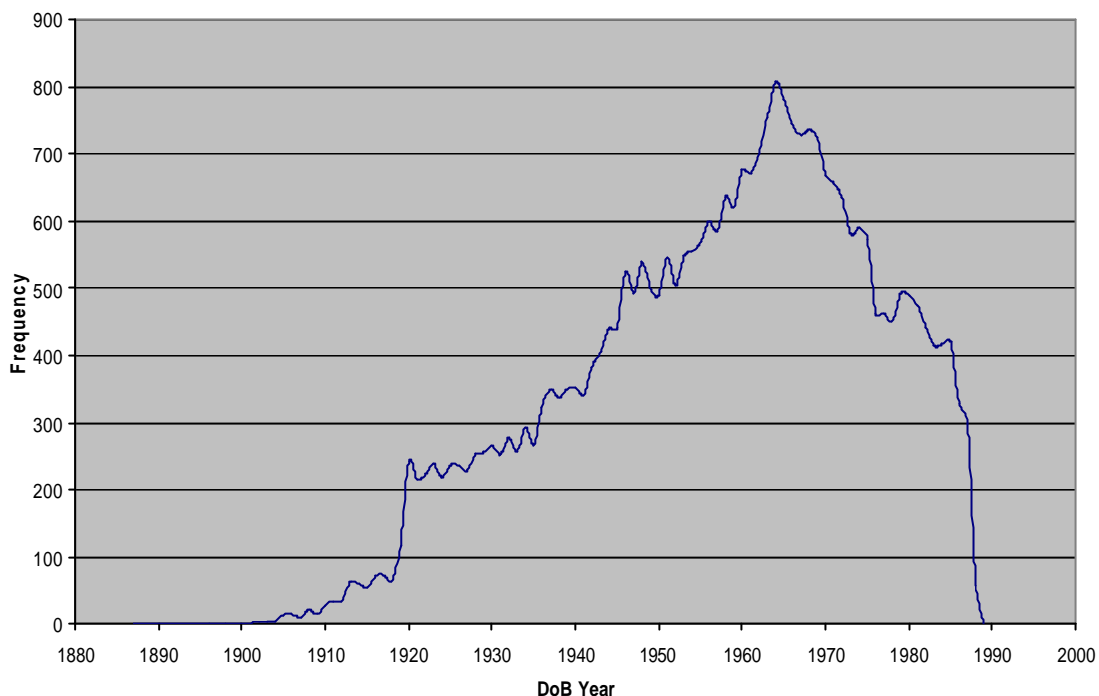
Value	Format	Frequency	Percentage s0
<b>not null</b>	<b>CCYYMMDD</b>	<b>39013</b>	<b>99.9</b>
0		21	0.1
840919		1	
	18??0000	3	
	19??0000	4	
19505000		1	
19545000		1	

3.2.2 The DoB range has been plotted by year in Figure 1. This has included demographics s1-s7 only (32370 records), to avoid spurious peaks caused by the date demographics. Corrections were made to the date value “840919” to make it 1984 (it is very unlikely to be 1884) and the 21 “zero” records were removed, leaving 32348 records included in the graph.

3.2.3 Note the high-value cut-off at 1988 (with just one record in 1989), which is to be expected: these people will be just 16 years old at the time of the data extract, the earliest age for which a licence can be issued. The peak occurs for 1964 with 808 records. The earliest four records are 1887, 1896, 1898, and two in 1901. There is a noticeable dip prior to 1920. These people would currently be approaching their 85<sup>th</sup> birthday, and in the absence of Date of Death information, the assumption is that only records of people currently living has been provided, perhaps explaining the dramatic reduction prior to 1920. Comparison with age statistics shows a broadly similar shape, although the 1946/7 baby boom is understated in the DVLA data.

3.2.4 There is no consistent evidence that default dates of birth occur for the decade – i.e. where the decade only is known. There is no peak for 1/1/1910, 1950, 1970,

1980, although there are noticeable peaks for 1920 and 1960, and small peaks for 1930 and 1940. Inspection of the data for these decades shows that only a handful of births occur on 1<sup>st</sup> January – too few to make any significant statistical interpretation.



**Figure 1 - DVLA DoB Frequency**

3.2.5 The same data (i.e. s1-s7) is plotted to show frequency of particular days in the year, to indicate if there is a particular peak at 1<sup>st</sup> Jan in any year. The year data was removed, the remaining month-day information sorted, and the occurrence of each of the 366 values counted and plotted in the graph below. (To simplify the process, the month-day data was combined with the year value 2004 (a leap year) to assist the production of the x-axis.)

3.2.6 Note the peak at 1<sup>st</sup> January and dip at 29<sup>th</sup> February. The average frequency for any given date is 88.3, and the standard deviation (sd) (ignoring data for 1/1 and 29/2) is 10 (11%) – nearly all the data falls within 1 sd, and most of it within 2. There is nothing else outside 3 sd. The theoretical average and sd for a probability of 1/365 (0.0027) and sample size as specified are 88.6 and 9.4 respectively – very close to the measured values.

3.2.7 The peak at 01/01 is 165, nearly twice the average value ( $165/88 = 1.88$ ), and some 8 sd larger. This is very unlikely to be random and we assume it is a real feature of the data, suggesting that approximately half of the 01/01 births are default values. The probability of a date being 01/01 is  $165 / 32370 = 0.0051$ . The theoretical sd for this value is 13 (i.e. 8%), and so in the full population, we might expect an occurrence rate of  $0.0051 \pm 0.0008$  (2sd) (i.e. within 16%). Combining the statistical uncertainties, we would expect an sd in this ratio of about 13%, i.e. 0.22. We do not know the full population size, but we do know the s8 and s9 sizes for the full population, i.e. 2485 and 4195 respectively. The

ratio s9/s8 should be similar to 1.88. In fact, taking into account the variation in birth rate between the two demographics, the ratio is derived as 2.1/1.1 or 1.98. The two values (1.88 and 1.98) are hence well within 1sd of each other, and therefore consistent.

3.2.8 The corollary is that  $1/1.88 = 53\% \pm 13\%$  (2sd) of 1<sup>st</sup> January dates are in fact correct, whereas 47%  $\pm 13\%$  are default dates.

3.2.9 The dip at 29/02 occurs because only 1 year in four is a leap year, and hence we would expect the number of occurrences of 29/02 to be approximately  $\frac{1}{4}$  of the average, i.e. 22. In fact there are 16 occurrences of 29/02, which is slightly lower than expected. However, the theoretical standard deviation for 22 occurrences out of 32,370 records is 4.7, so the measured value 16 is within 2sd, so is consistent.

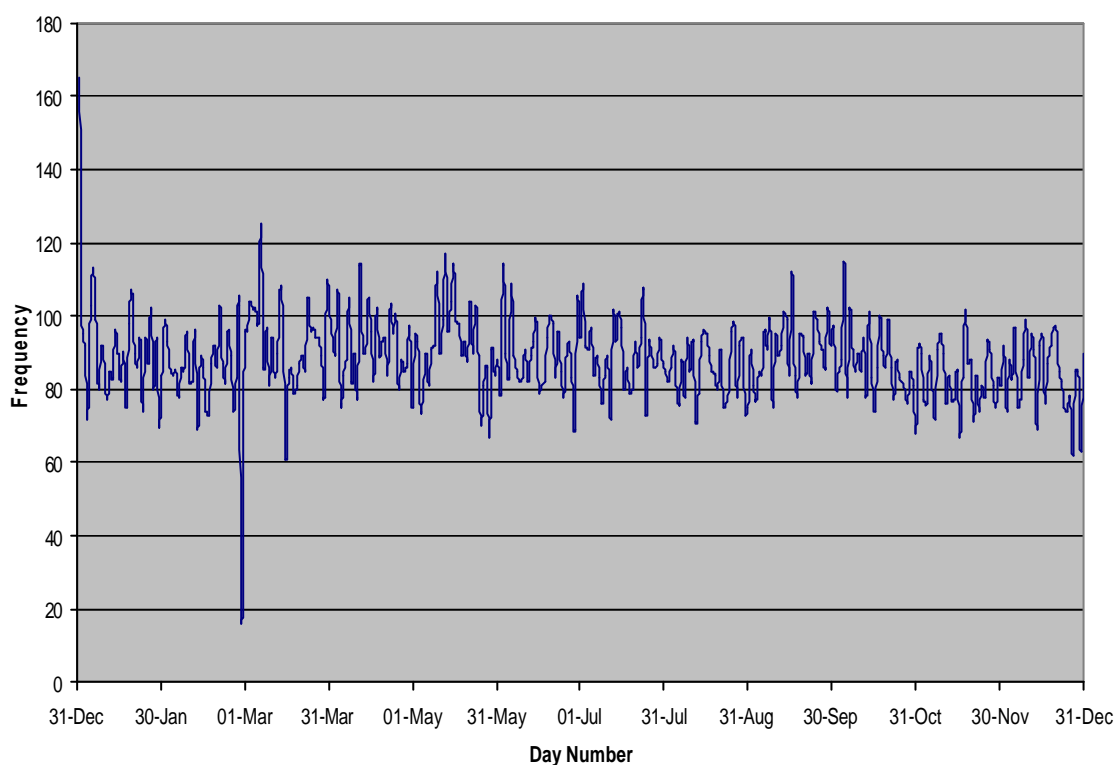


Figure 2 - DVLA DoB by Day Number

### 3.3 Gender

Value	Format	Frequency	Percentage s0
Female	F	16658	42.7
Male	M	22347	57.3
Null		7	0
Invalid	1	1	0

### 3.4 Place of birth

3.4.1 Shows town or country of birth. Over 15000 records show a country not a town.

Value	Format	Frequency	Percentage s0
<b>Null</b>		<b>9089</b>	<b>23.3</b>
<b>Not null</b>		<b>29924</b>	<b>76.7</b>
England		7513	19.3
Not Known	NK	3774	9.7
United Kingdom		1959	5.0
London		1219	3.1
Wales		862	2.2
Pakistan		636	1.6

3.4.2 Title

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
undefined	0	7	0
Mr	1	22129	56.7
Mrs	2	9425	24.2
Miss	3	6283	16.1
Female – No Prefix Title	4	66	0.2
Male – No Prefix Title	5	7	0
Female – Prefix Title	6	884	2.3
Male – Prefix Title	7	212	0.5
Female – Title contains full mode of address	8	0	0
Male – Title contains full mode of address	9	0	0

### 3.5 Name elements

#### Surname

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>39013</b>	<b>100</b>

#### Forename(s)

3.5.1 Contains all forenames of a person.

3.5.2 There were 600+ records which include an asterisk suffix (e.g. Benjamin\*). The reason for this is unclear.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>39013</b>	<b>100</b>
	Includes * suffix	>600	>1.5
Initial		13	0
Dash		2	0

### Alternative Surname

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>32478</b>	<b>83.2</b>
<b>not null</b>		<b>6535</b>	<b>16.8</b>

### Alternative Forename(s)

#### 3.5.3 All alternative forenames

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>32478</b>	<b>83.2</b>
<b>not null</b>		<b>6535</b>	<b>16.8</b>
Asterisked (*)		>140	>0.2

## 3.6 Address lines

### Address line 1

#### 3.6.1 First line of address commonly containing house or flat name or number.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>7</b>	<b>0</b>
<b>not null</b>		<b>39006</b>	<b>100</b>

### Address line 2

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>18893</b>	<b>48.4</b>
<b>not null</b>		<b>20120</b>	<b>51.6</b>

### Address line 3

#### 3.6.2 Four records have dashes followed by either CASP CASE or CASP NUMBER and then a number.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>36790</b>	<b>94.3</b>
<b>not null</b>		<b>2223</b>	<b>5.7</b>
CASP -----		4	0

### Address line 4

Value	Format	Frequency	Percentage s0
-------	--------	-----------	---------------

<b>null</b>		<b>38928</b>	<b>99.2</b>
<b>not null</b>		<b>85</b>	<b>0.2</b>

### Address line 5

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>36749</b>	<b>94.2</b>
<b>not null</b>		<b>2257</b>	<b>5.8</b>

### Postcode

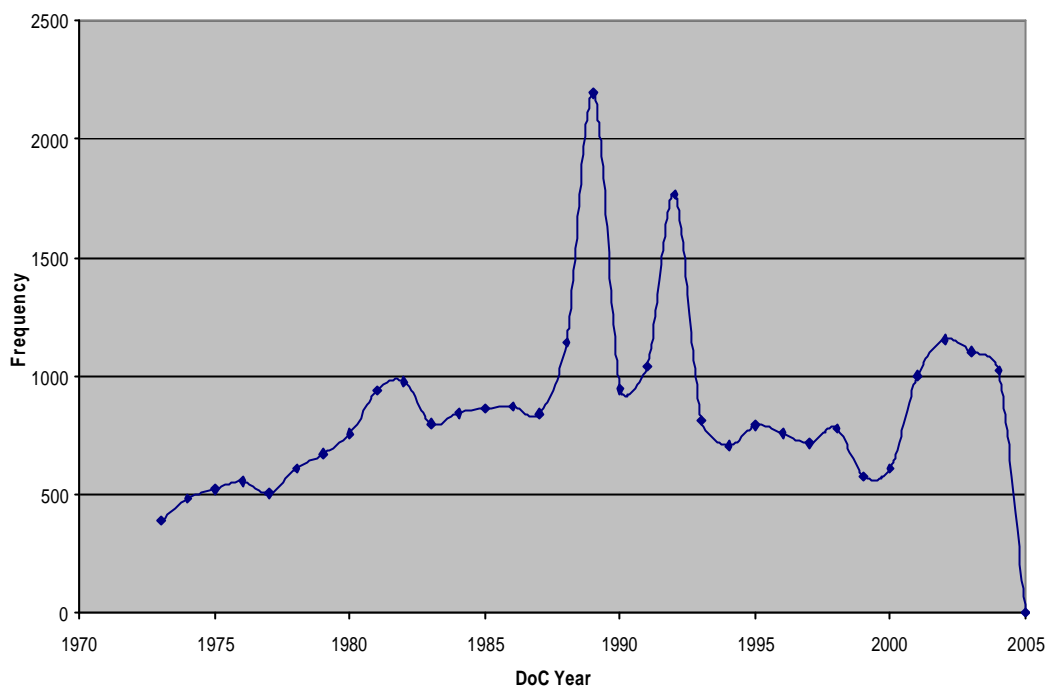
- 3.6.3 DVLA has dummy postcodes AA88 and AA89. Postcodes of AA88 are for postal towns spelt in Welsh postcodes of AA89 show changed postal towns as described above.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>7</b>	<b>0</b>
<b>not null</b>		<b>39006</b>	<b>100</b>
AA88		12	0
AA89		79	0.2

## 3.7 Account creation date

- 3.7.1 First date of provisional licence issue.
- 3.7.2 Two dates appear invalid as shown below.

Value	Format	Frequency	Percentage s0
<b>Zero</b>		<b>11236</b>	<b>28.8</b>
<b>Non Zero</b>	<b>CCYYMMDD</b>	<b>27777</b>	<b>71.2</b>
19200475		1	0
970710		1	0



**Figure 3 - DVLA Date of Creation**

3.7.3 Ignoring the spurious datum for 1920, the data ranges from 1973 through to 2005 – just three records from the first few days of 2005 have made it in. The data is broadly flat with an average of 841. An initial growth from 1973 to 1982 was followed by a flatter period through to c. 2000 with two noticeable peaks in 1989 (2195) and 1992 (1768). There also appears to be a slight increase in the recent period 2001-2004 with c. 1050 records pa.

### 3.8 Last update date

3.8.1 The initial DVLA data extracted missed out the Last Update column. This column is of particular interest to the trial, in order to understand the currency of the DVLA data. Subsequent to the completion of the main part of Lot 1 as published in version 1.0 of this document, a new extract file from DVLA has been analysed, which includes the Last Update column. The results for this are shown below.

Value	Format	Frequency	Percentage s0
Zero		8696	22.2%
Text (garbage)		1234	3.1%
Non Zero	CCYYMMDD	29266	74.7%

3.8.2 Note the significant amount of garbage data in this column. A large number (c. 1300) of ASCII NUL characters was also observed embedded in some of the post code fields (apparently 6-character post-code fields that were padded with a trailing NUL to pad out the field). Note that these were not present in the original extract, where this padding character was a space. As an aside, the issue with the original extract files, where the ASCII SOH character occurred in every row, has gone away in the new extract. Because of these differences between the new and original extracts, we speculate that there is a link between how the

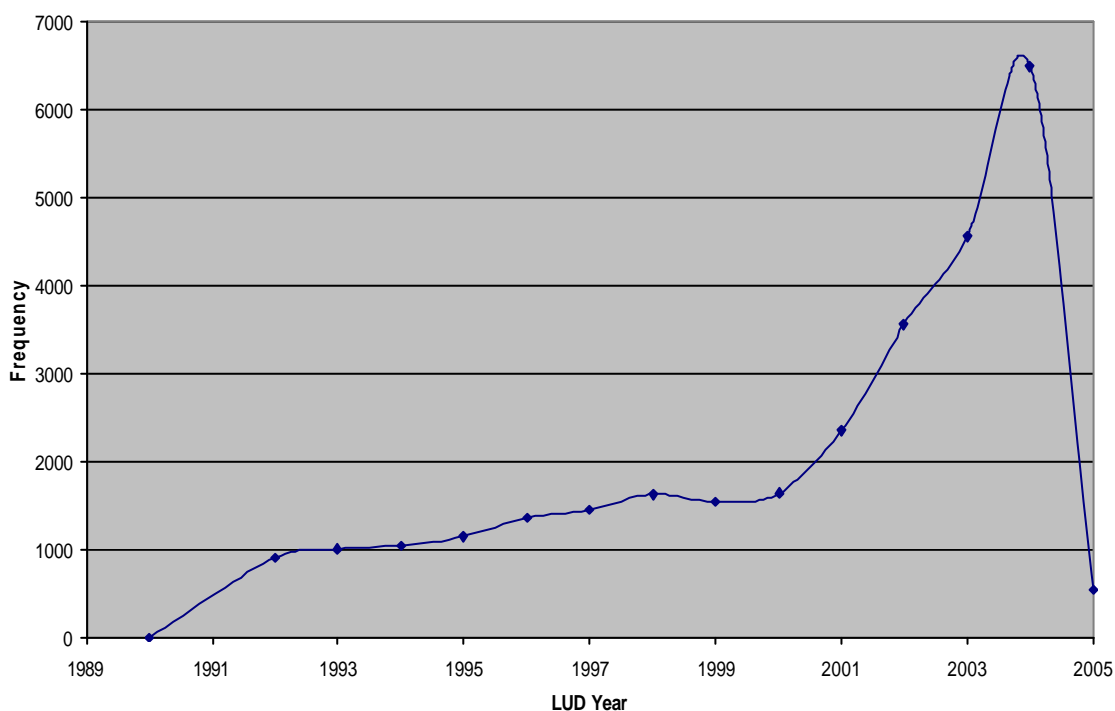
extract program handles NULs (and other ASCII control codes), and the appearance of garbage data in the Last Update Date column. In other words, we suspect that the garbage data is not actually present in the DVLA database Last Update Date, but is an erroneous artefact of the extract process.

3.8.3 A sanity check was also carried out by analysing the Date of Creation field. This was plotted and shown to have an almost identical format as before (the new graph is not included in this report). The new extract had 39196 records, some 183 records more than before. Brief analysis of the DoC field indicated that compared to the original dataset, the new dataset had:

- 81 more records for year 2005, to be expected.
- 50 more records for year 2004, not unexpected
- a few records more or less for all other years, possibly indicating name or address changes that have taken records out or brought records into the demographic extract.

3.8.4 We therefore conclude that the new extract file is close enough to the old to warrant comparison.

3.8.5 The valid Last Update Date data has been analysed, and is presented in Figure 4 - DVLA Last Update Date.

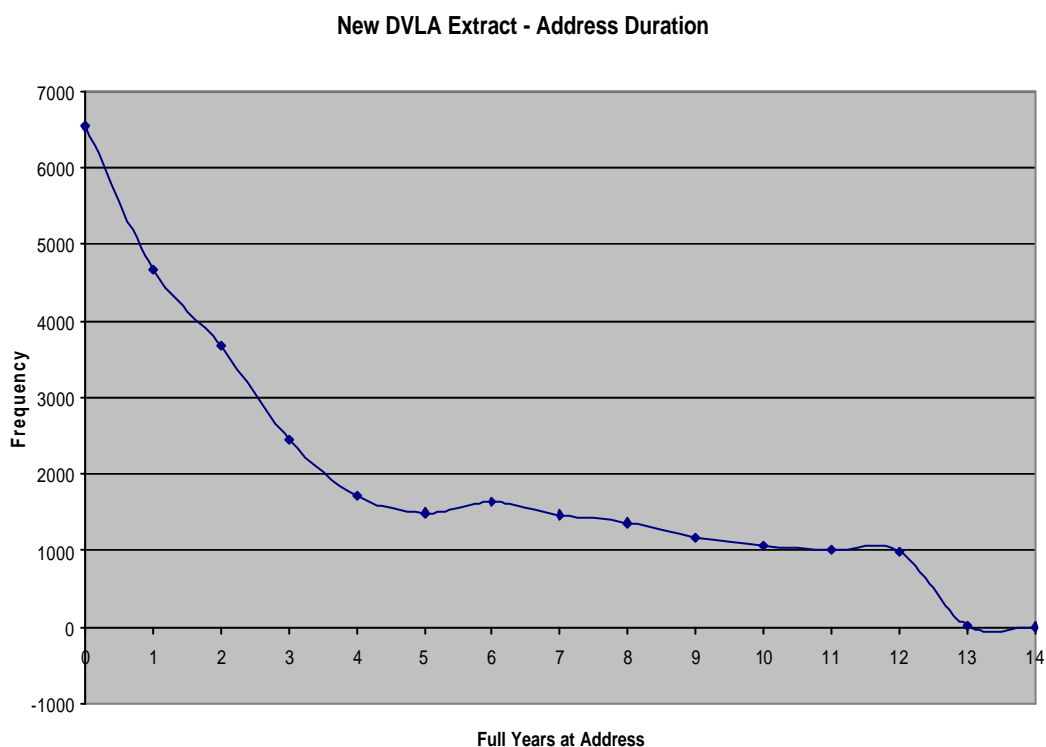


**Figure 4 - DVLA Last Update Date**

3.8.6 The Last Update Date ranges from 1974 to now. Only a few records are shown for 2005, which is to be expected. Few records exist pre 1991, and so these have been left off the graph. The frequencies are tabulated below.

LUD Year	Frequency
1974	1
1975	1
1978	2
1979	2
1982	4
1984	2
1986	1
1987	3
1988	2
1989	2
1990	2

3.8.7 For comparison with HMRC address durations, the above figure is also plotted as a duration, in Figure 5.



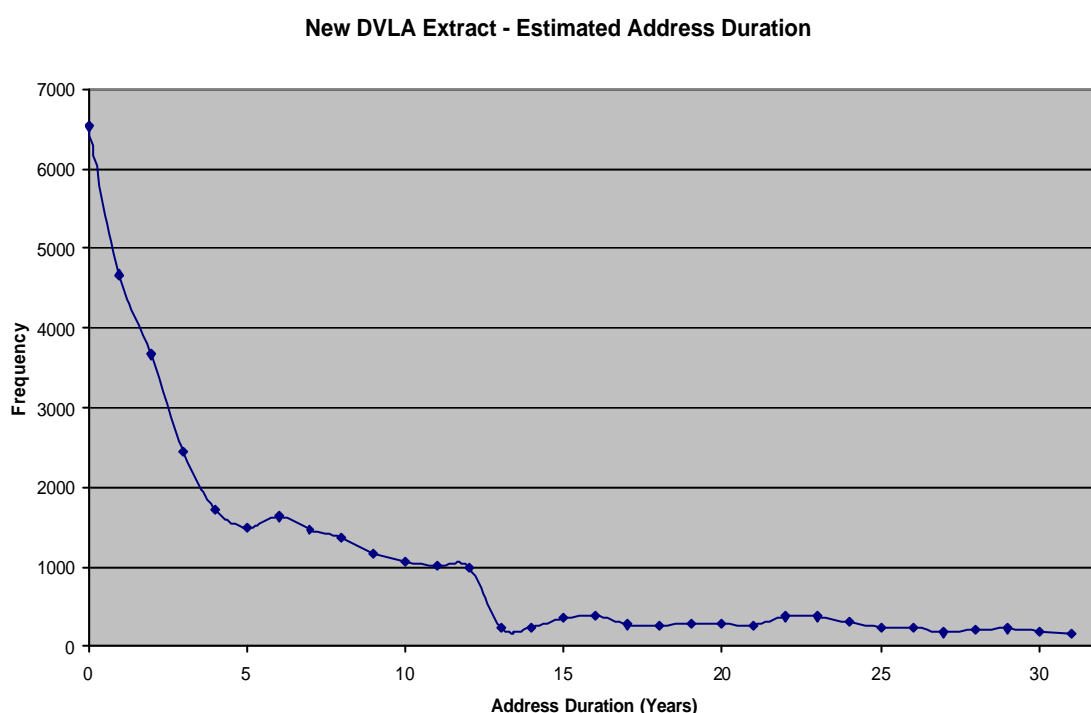
**Figure 5 - DVLA Address Duration**

3.8.8 Once again, the handful of addresses occupied for more than 14 years has been omitted from the graph for clarity. The largest peak is for people who have changed address in the last 12 months, and it falls steadily. Nearly all addresses have been occupied for less than 12 years. The sharp drop at 12 years is not seen in the HMRC data, so is probably not representative of people’s behaviour nor of particular events 13 years ago (1991). We are therefore suspicious that this may be an anomaly with the DVLA data.

3.8.9 Note that there are c. 10,000 records that have a zero or garbage value, and this may explain the missing data. It is likely that a person who has not moved since applying for a driving licence will have a Last Update Date of zero. Hence, for these records, we can take the Creation Date as the “last update date”. In that case, we find the following:

Value	Frequency	Percentage s0
DoC used where LUD is zero	4993	12.7%
LUD and DoC both zero	3705	9.5%

3.8.10 In other words, there are an additional 4993 records we can add to the graph, and these have been plotted in Figure 6.



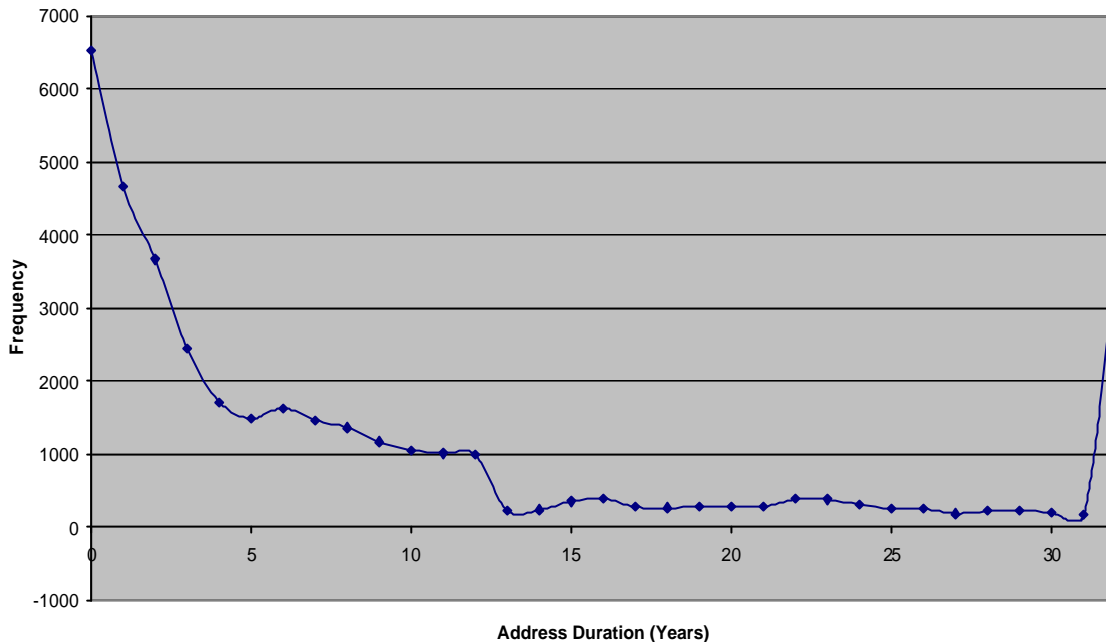
**Figure 6 - DVLA Address Duration 2**

3.8.11 There is still a dip at duration 13 years, inconsistent with HMRC; but we do now see data consistent with HMRC for durations above year 15. In fact, the data extends up to 31 years, consistent with the oldest DoC value of 1973.

3.8.12 A third analysis has been carried out, this time attempting to account for the records where both Last Update Date and Date of Creation are null. These appear to be for people whose 17<sup>th</sup> birthday is prior to 1973. We might assume that these people have never changed their address since they turned 17, in which case they have been living at the same address for more than 31 years (in some case up to 80 years). These addresses have therefore been added to the year-32 “bucket”, as shown in Figure 7 by the peak in year 32, similar to the peak seen in the HMRC data. **WARNING:** we do not recommend this third analysis as a safe interpretation of the address duration data. While it is likely to be correct in

some cases, an alternative explanation for other cases is that the drivers concerned have failed to update DVLA with their current address.

**New DVLA Extract - Address Duration 3**



**Figure 7 - DVLA Address Duration 3**

3.8.13 The average time at an address has been calculated from the graph data as follows:

- Duration 1 4.2 years
- Duration 2 6.8 years
- Duration 3 10.8 years

3.8.14 These should be compared with the HMRC “current duration” of c. between 4 and 9 years.

## 4. Identity duplication

### 4.1 Date of birth, name and address matching

- 4.1.1 Occurrence of duplicate records for DVLA is 0.14% or 64 records out of the 45,539 split records.
- 4.1.2 There are 64 families, with 64 parents and 64 members, i.e. each family contains just two records. These appear to be genuine cases of an individual holding two DVLA identities, i.e. with two driver numbers. The total number of driver records is 39,013, so the underlying duplication rate is 64 people out of (39,013 – 64) people, or 0.17%.
- 4.1.3 Results post QAS address cleansing are very similar, with 64 parents and 65 members. The number of family groups is the same, but one additional member record was matched resulting in one family group having two members (i.e. the individual concerned has three driver numbers). In fact, closer analysis of the differences shows the following:

#### Pre QAS

- 4.1.4 The following (anonymised) matched pair of records exists

- Mr Barry Michael Smyth                      48 Hollow Valley Block, Bristol BS38RU
- Mr Barry Michael Smith                      48 Hollowvalley Block, Bristol BS38RD

- 4.1.5 Post QAS, the Mr Smith record was corrected to be “Flat 48”, but the Mr Smyth record, with a post code that is valid but wrong, failed to match in QAS with sufficient confidence to be accepted, and hence remained as was. As a result, there were too many differences for the records to match. A manual search within the QAS Address with Names data indicates that indeed a Mr Smyth lives at 48 Hollowvalley Block, and it would appear that the match should be counted.

#### Post QAS

- 4.1.6 As mentioned, the above Mr Smyth pair of records failed to match. However, an additional pair of records, whose raw (pre-QAS) values were as follows,

- Mr Mustafa                                      14 Oversetts Street, Derby      DE238HR
- Mr Mustafa                                      14 Augusta Street, Derby        AA89

- 4.1.7 These were both corrected by QAS to 14 Augusta Street, DE23 8HR, and now did succeed in matching.

4.1.8 Finally, another AA89 postcode record was successfully QAS corrected and joined the existing family as follows:

- MR FALIR /FADK            21 STOCKTON AVENUE        BS96PD
- MR FALAR /FADIK        21 STOUGHTON AVENUE    BS9 6BD
- MR FALA /FADIC        21 STOUGHTON AVENUE    BS9 6BD

4.1.9 This particular set of records is interesting in that there are a variety of spelling differences in the name data, and subtle differences in the address data, perhaps suggesting deliberate obfuscation of the identity (i.e. fraud). Note the differences in name, both surname and forename, and in street and postcode. Note that the 21 STOCKTON AVENUE address is not a valid postal address (which is why the postcode format has not been corrected by QAS). Also, in the raw input, the FADIK record had the incorrect postal code of BRISTOL AA89, which was corrected by QAS to BS9 6BD, and that is why it then appeared in the post QAS match.

4.1.10 In summary, the actual number of records should therefore be 65 families and 66 members, in other words:

4.1.11 Total number of individual people records is  $(39,004 - 66) = 38,938$ , and the number of people with more than one record is 65 or 0.17%. There is one person with three records.

4.1.12 Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of:

$$0.17\% \pm 2sd = \\ 0.17\% \pm 0.04\%$$

4.1.13 In a database of 30million drivers, this equates to between 39,000 and 63,000 duplicates.

4.1.14 Note that there are potentially additional records that failed to have the address cleansed, that would have otherwise matched. Hence the true match rate within the sample is likely to be even larger.

## Appendix B: Data assessment - GRO

---

## 1. Analysis of datasets

1.1.1 The GRO data sample contains 8,700 birth records and 4,272 death records. There are 3 birth records that were sampled twice via the various demographics.

### 1.2 Coverage

1.2.1 Analysis of this dataset shows the following:

- Taken together, GRO and GROS coverage appears to be approximately 20% of the (UK) population for births and only 13% for deaths.
- The GRO data covers a period from 1993. Assuming an average age of 75 years we would expect coverage of about 12/75ths or 16% of the (England and Wales) population. For Births, this covers a current age range of 0 to 12 years.
- The Birth and Death coverage for s6 Wales is only half that of the other demographics. This suggests an absence of both older people and young families in this location, which is contrary to expectation.
- The low death rate for s7 Birmingham also suggests an absence of the older generation. This is possible in a high-density urban area, but not expected if low-income is associated with both high-density housing and the older generation.

### 1.3 Field analysis

1.3.1 The data comprised the following key fields:

- Unique reference number
- Date of Birth
- Date of Death
- Gender
- Place of Birth
- Surname and separate Forename fields
- An Address and Postcode fields
- Account Creation Date
- Second name (45% populated in Births)
- Third name (aliases) (4% populated in Deaths)
- Second address (2% populated in Births)
- Third address (11% populated in Births)

## 2. Data structure

2.1.1 The GRO data covers England and Wales only, and was supplied as two separate datasets, one containing births and the other deaths data. The

demographics s1-s4 and s6-s7 only were supplied. The demographic s5 was omitted because it is of Scottish data and included in the GROS data-set and the two birthdates selected in s8 and s9 predate the GRO database.

2.1.2 The file format was well-formed csv with quotes around data values.

### 3. Statistical summary

#### GRO – s1 – Typical Dataset by name

3.1.1 Unlike other samples the matching criteria has included hyphenated or double barrelled name of which the criteria applies to the second half of the name including names with a prefix of Mc where the first four characters match the criteria. E.g. if the criteria was DON\*, then the names Donald, McDonald and Cooper-Donald are all seen in the sample.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>728</b>	<b>8.4</b>
Criteria not met in surname (total): Made up of:-	188	
Criteria only met in Surname 2 (AKA)	169	
Criteria not met at all	4	
Criteria only met in second half of name (hyphenated or space)	13	
Criteria met in second half of Surname 2 (hyphenated or space)	2	
<b>Deaths</b>	<b>441</b>	<b>9.6</b>

#### GRO – s2 – Typical Dataset by name

3.1.2 Surname based sample with similar finding to above, where the latter part of hyphenated or double barrelled surnames match the specified criteria. No Mc prefixes occurred in this demographic, probably because there are no Mc-names matching the particular name stem specified.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>623</b>	<b>7.2</b>
Criteria not met in surname (total): Made up of:-	150	
Criteria only met in Surname 2	123	
Criteria not met at all	8	
Criteria only met in second half of name (hyphenated or space)	16	
Criteria met in second half of Surname 2 (hyphenated or space)	3	
<b>Deaths</b>	<b>396</b>	<b>9.3</b>

**GRO – s3 – Typical suburban dataset by geographic area (postcode and area name)**

The area name criteria was met in many records which did not have the specified postcode, this was common with other data-sources.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>1898</b>	<b>21.8</b>
<b>Deaths</b>	<b>1706</b>	<b>39.9</b>

**GRO – s4 – Covers name issues and address issues on houses that have been converted into flats. (postcode)**

Postcodes were of correct format and met criteria in every record.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>2733</b>	<b>31.4</b>
<b>Deaths</b>	<b>590</b>	<b>13.8</b>

**GRO – s6 – Covers issues around Welsh names and addresses (postcode and area name)**

One record match the area name but was well outside of intended area.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>472</b>	<b>5.4</b>
<b>Deaths</b>	<b>348</b>	<b>8.2</b>
Match outside area	1	

**GRO – s7 – Covers issues related to high density urban areas and high rise flat blocks**

Postcodes were of correct format and met criteria in every record.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>2246</b>	<b>25.8</b>
<b>Deaths</b>	<b>791</b>	<b>18.5</b>

## 4. Individual data item analysis

### 4.1 Unique ID

Unique identifier

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>distinct</b>		<b>8700</b>	<b>100</b>
	<b>null</b>		<b>0</b>	<b>0</b>
<b>Death</b>	<b>distinct</b>		<b>4272</b>	<b>100</b>
	<b>null</b>		<b>0</b>	<b>0</b>

This not null unique identifier shows 3 people intersecting two demographics.

Intersected Demographic	Frequency
<b>Birth</b>	
S2-s4	2
S2-s7	1

### 4.2 Reference number

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>distinct</b>		<b>8</b>	
	<b>null</b>		<b>8589</b>	
	<b>duplicate</b>		<b>2</b>	
<b>Death</b>	<b>distinct</b>		<b>8</b>	
	<b>null</b>		<b>4264</b>	

#### Startdatetime of reference number (date valid from)

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>distinct</b>	<b>CCYYMMDD</b>	<b>8700</b>	<b>100</b>
	<b>null</b>		<b>0</b>	<b>0</b>
<b>Death</b>	<b>distinct</b>	<b>CCYYMMDD</b>	<b>4272</b>	<b>100</b>
	<b>null</b>		<b>0</b>	<b>0</b>

#### Enddatetime of reference number (date valid to)

Should be 99999999.

	Value	Format	Frequency	% s0_B or s0_D
--	-------	--------	-----------	----------------

<b>Birth</b>	999999999	CCYYMMDD	8700	100
	null		0	0
<b>Death</b>	999999999	CCYYMMDD	4272	100
	null		0	0

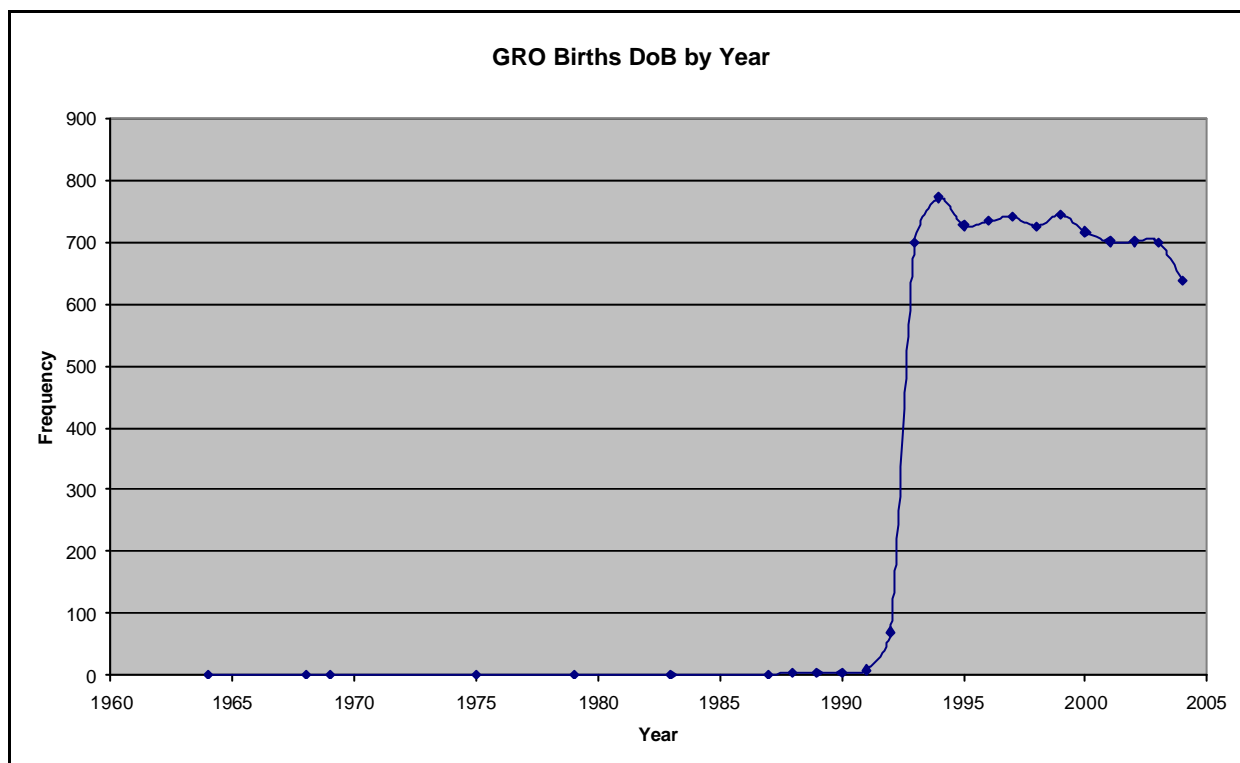
**Temp Reference Number, Startdatetime of temp ref no, Enddatetime of temp ref no, Verified ID**

All the above fields are null in all recrds.

### 4.3 GRO births – Date of birth

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Distinct</b>	<b>CCYYMMDD</b>	<b>8700</b>	<b>100</b>
	<b>Null</b>		<b>0</b>	

GRO Births DoB ranges from 1964 to 2004 as shown in Figure 8, below



**Figure 8 – GRO Births DoB by Year**

4.3.1 Note the single values for the years 1964 to 1983 recorded in the table below and also just visible on the graph. These are unexpected, as the Account Creation Date field indicates that the database was not started until 1993. Otherwise, the graph is expected, with a nominally flat (reducing?) birth rate since 1994.

Year Value	Frequency
1964	1

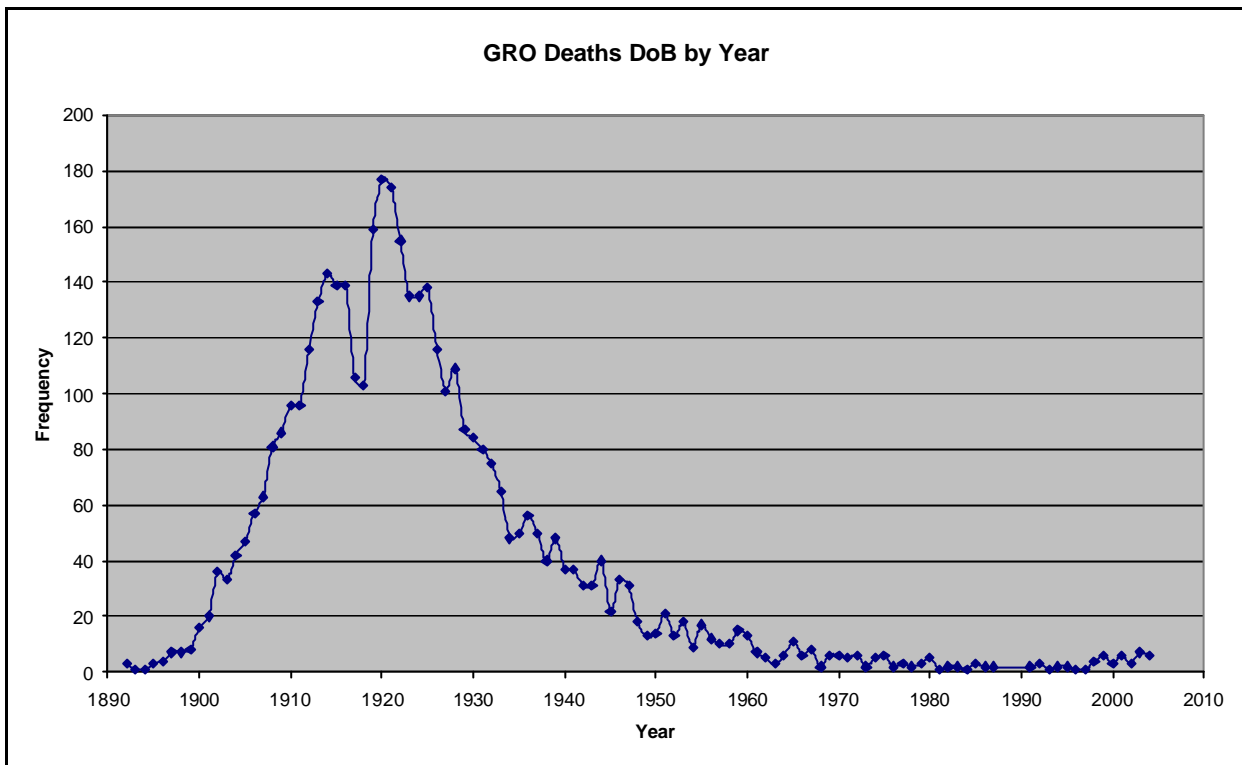
1968	1
1969	1
1975	1
1979	1
1983	1
1987	2
1988	4
1989	5
1990	4
1991	9
1992	70

Etc

#### 4.4 GRO deaths – Date of birth

<b>Death</b>	<b>Distinct</b>	<b>CCYYMMDD</b>	<b>4272</b>	<b>100</b>
	<b>Null</b>		<b>0</b>	
		CCYY0000	4	
	19120600		1	
	19200200		1	

4.4.1 The range for Deaths DoB extends back to 1892 and is shown in the following graph. As expected, the Date of Birth data for GRO Deaths is weighted towards earlier dates indicating that it is mostly older people who die. It shows a marked peak around 1920 – indicating a median life expectancy of 70 to 80 years. The pronounced dip occurs at 1917 and 1918 – consistent with the dip in birth rate during WW1 (see section **Error! Reference source not found., Error! Reference source not found.**).



**Figure 9 – GRO Deaths: DoB by Year**

4.4.2 An inspection of the DoB data indicates that there is an occurrence of “default” dates, where only the birth year has been recorded, or a month but zero for the day. The following figure shows the frequency of particular dates, ignoring the invalid dates with a day value of zero. There is no apparent default of 1<sup>st</sup> Jan or 1<sup>st</sup> of any other month. The obvious dip is for 29<sup>th</sup> February as expected.

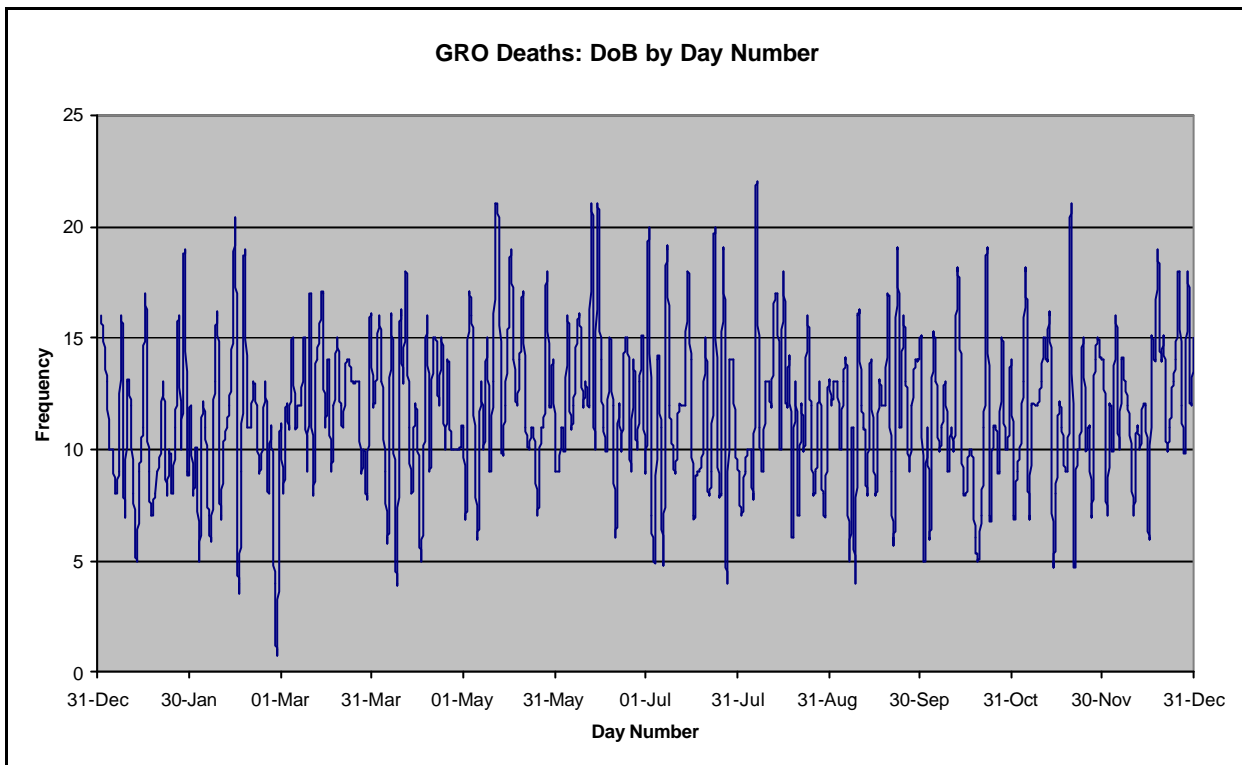
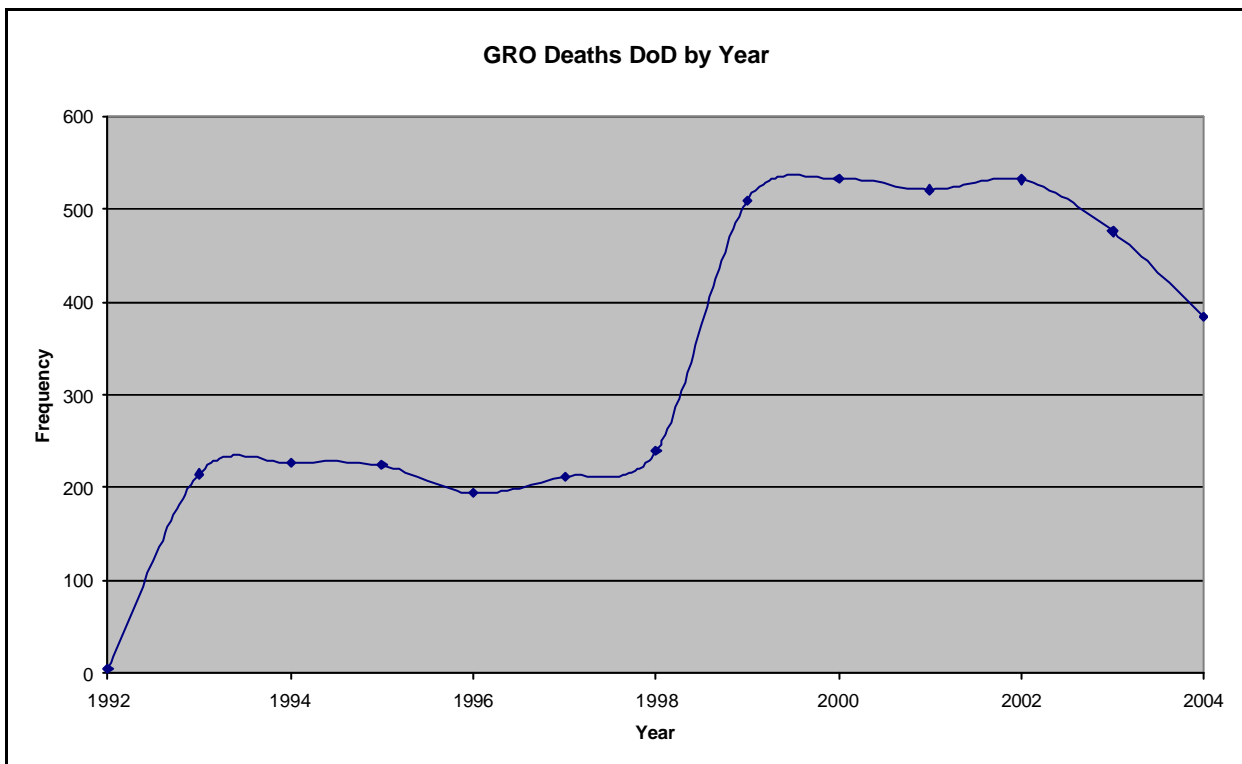


Figure 10 – GRO Deaths: DoB by Day Number



#### 4.5 Verified date of birth

Null in all records.

## 4.6 Date of death

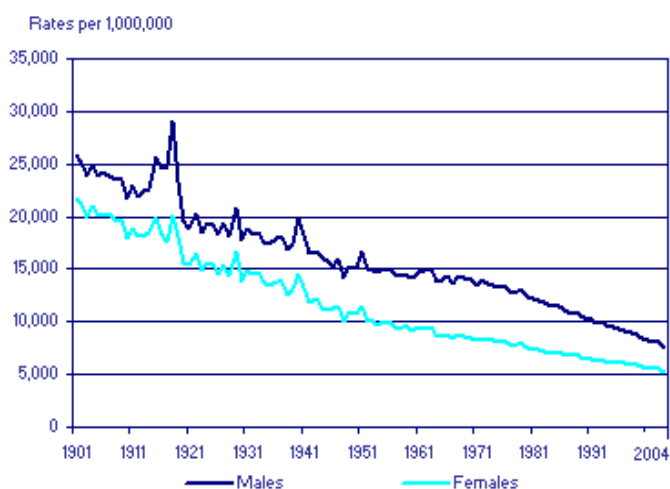
Only in death data.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>distinct</b>	<b>CCYYMMDD</b>	<b>0</b>	
	<b>null</b>		<b>8700</b>	<b>100</b>
<b>Death</b>	<b>distinct</b>	<b>CCYYMMDD</b>	<b>4272</b>	<b>100</b>
	<b>null</b>		<b>0</b>	

4.6.1 DoD ranges from 1992 (5 records) to date, as is shown in the following graph. Note that the overall shape of the graph is similar to the Account Creation Date shown in Figure 16, which is to be expected.

**Figure 11 – GRO Deaths: DoD by year**

4.6.2 The death rate shows an unexpected rise in 1999, with a death rate twice that seen for previous years. The reason for this is not clear. For comparison, Figure 12 shows the UK national death rates for the years 1901 to 2003. Although the graph is small, the last portion covering 1993 to 2003 can be compared with Figure 11. In contrast, the national statistics seem to show the opposite trend – a higher rate followed by a drop from year 2000 onwards. We do not know the reason for this difference.



**Figure 12 – Death Statistics for UK (source ONS)**

4.6.3 A view of life expectancy can be found from the following figure which shows age of the deceased, determined from the difference (DoD – DoB). The average age is 75.2. Note the high infant mortality with 37 deaths under one year. Of these, 21 occurred within the first 3 months.

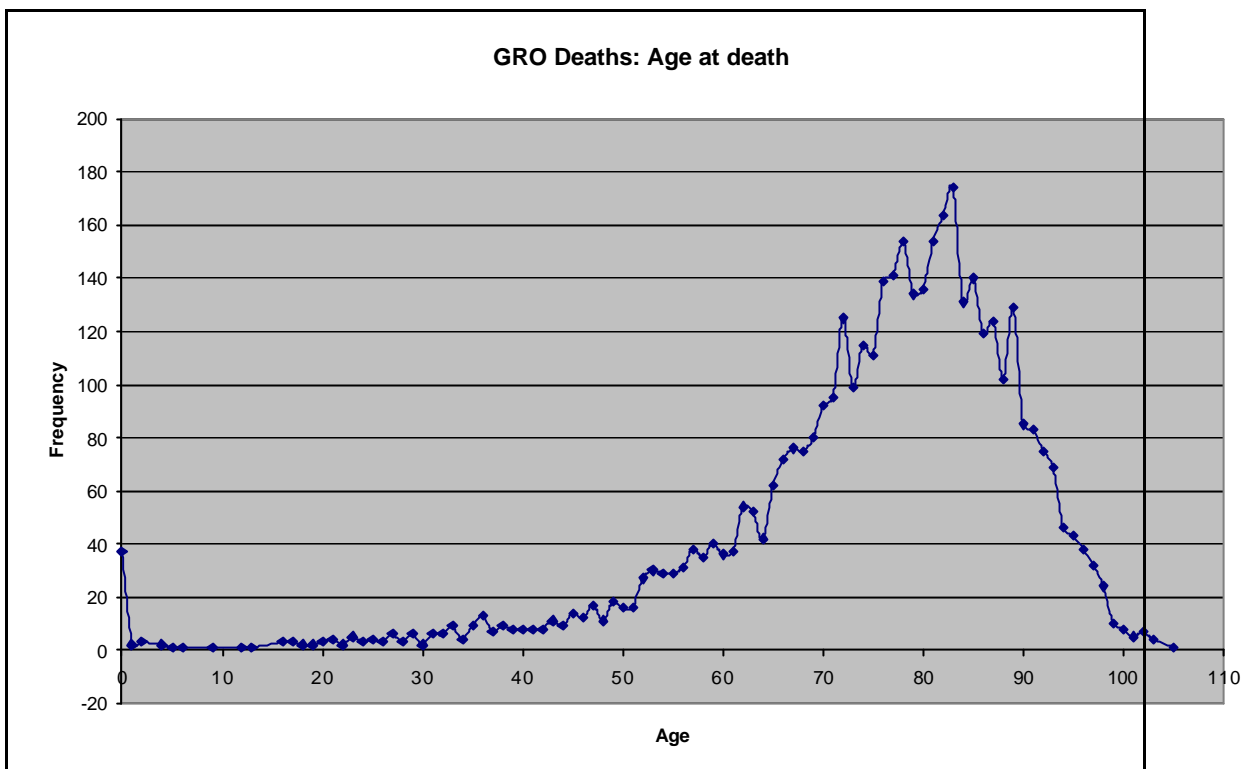


Figure 13 – GRO Deaths: Age at Death by year

## 4.7 Gender

### Gender (Sex as received from RSS or paper copy)

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Female</b>	<b>F</b>	<b>4215</b>	<b>51.6</b>
	<b>Male</b>	<b>M</b>	<b>4485</b>	<b>48.4</b>
	<b>null</b>		<b>0</b>	
<b>Death</b>	<b>Female</b>	<b>F</b>	<b>2172</b>	<b>50.8</b>
	<b>Male</b>	<b>M</b>	<b>2100</b>	<b>49.2</b>
	<b>null</b>		<b>0</b>	

### Gender (Sex text from paper copy)

Null in all records.

## 4.8 Place of birth

4.8.1 Address data of place of birth, most frequently in format of hospital name, town for births and town or country for deaths

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>0</b>	<b>0</b>
	<b>not null</b>		<b>8700</b>	<b>100</b>
	Hospital, Town 1		1023	11.8
	Hospital, Town 2		822	9.4

	Hospital, Town3		800	9.2
	Other values		6055	69.6
<b>Death</b>	<b>null</b>		<b>10</b>	<b>0.4</b>
	<b>not null</b>		<b>2446</b>	<b>99.6</b>
	Town 1		463	10.8
	Town 2		148	3.5
	London		134	3.1
	Town 2 County		112	2.6
	Irish Republic		104	2.4
	Town 2 County		71	1.7
	India		69	1.6
	Dashed line (various)		8	0.2
	Other values		3153	73.8

## 4.9 Name elements

### Name 1 Surname

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>1</b>	<b>0</b>
	<b>not null</b>		<b>8699</b>	<b>100</b>
<b>Death</b>	<b>null</b>		<b>0</b>	<b>0</b>
	<b>not null</b>		<b>2456</b>	<b>100</b>

### Name 1 Forename 1 (First forename)

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>2</b>	<b>0</b>
	<b>not null</b>		<b>2733</b>	<b>100</b>
	Dashed line		2	
<b>Death</b>	<b>null</b>		<b>0</b>	<b>0</b>
	<b>not null</b>		<b>2456</b>	<b>100</b>
	Dashed line		1	

### Name 1 Forename 2

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>1044</b>	<b>12</b>
	<b>not null</b>		<b>7656</b>	<b>88</b>
	Dashed line		1	
<b>Death</b>	<b>null</b>		<b>1085</b>	<b>25.4</b>
	<b>not null</b>		<b>3187</b>	<b>74.6</b>
	Dashed line		1	

### Name 1 Forename 3

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>7072</b>	<b>81.3</b>

	<b>not null</b>		<b>1628</b>	<b>18.7</b>
	Dashed line		1	
<b>Death</b>	<b>Null</b>		<b>3886</b>	<b>91</b>
	<b>not null</b>		<b>386</b>	<b>9</b>

### Name 1 Additional Forename(s) - Any additional forenames.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8530</b>	<b>98.1</b>
	<b>not null</b>		<b>170</b>	<b>1.9</b>
<b>Death</b>	<b>Null</b>		<b>4248</b>	<b>99.4</b>
	<b>not null</b>		<b>24</b>	<b>0.6</b>

### Name Title

Null in all records

### Name 2 Surname - Surname of AKA name.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>4810</b>	<b>55.3</b>
	<b>not null</b>		<b>3890</b>	<b>44.7</b>
	Dashed line		1	0
<b>Death</b>	<b>Null</b>		<b>4051</b>	<b>94.8</b>
	<b>not null</b>		<b>221</b>	<b>5.2</b>

### Name 2 Forename 1

First forename.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>4811</b>	<b>55.9</b>
	<b>not null</b>		<b>3889</b>	<b>44.1</b>
	Dashed line		2	
<b>Death</b>	<b>Null</b>		<b>4051</b>	<b>94.8</b>
	<b>not null</b>		<b>221</b>	<b>5.2</b>

### Name 2 Forename 2

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>5189</b>	<b>59.6</b>
	<b>not null</b>		<b>3511</b>	<b>40.4</b>
	Dashed line		1	
<b>Death</b>	<b>Null</b>		<b>4156</b>	<b>97.3</b>
	<b>not null</b>		<b>116</b>	<b>2.7</b>

### Name 2 Forename 3

	Value	Format	Frequency	% s0_B or s0_D
--	-------	--------	-----------	----------------

<b>Birth</b>	<b>Null</b>		<b>7885</b>	<b>90.6</b>
	<b>not null</b>		<b>815</b>	<b>9.4</b>
<b>Death</b>	<b>Null</b>		<b>4255</b>	<b>99.6</b>
	<b>not null</b>		<b>17</b>	<b>0.4</b>

### Name 2 Additional Forename Initial

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8622</b>	<b>99.1</b>
	<b>not null</b>		<b>78</b>	<b>0.9</b>
<b>Death</b>	<b>Null</b>		<b>4270</b>	<b>99.9</b>
	<b>not null</b>		<b>2</b>	<b>0.1</b>

### Name 2 Title

Null in all records

### Aliases

Always null for birth records.

Mostly prefixed with “formerly known as”.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8700</b>	<b>100</b>
	<b>not null</b>		<b>0</b>	<b>0</b>
<b>Death</b>	<b>Null</b>		<b>4291</b>	<b>95.8</b>
	<b>not null</b>		<b>181</b>	<b>4.2</b>

## 4.10 Address elements

### Address 1

Full Address of mother for births data or of the deceased for death data. Often nursing home for death data.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>174</b>	<b>2</b>
	<b>not null</b>		<b>8526</b>	<b>98</b>
<b>Death</b>	<b>Null</b>		<b>2</b>	<b>0.1</b>
	<b>not null</b>		<b>4270</b>	<b>99.9</b>
	Dashed line		<b>1</b>	

### Postcode

Postcode data was of valid format.

	Value	Format	Frequency	% s0_B or s0_D
--	-------	--------	-----------	----------------

<b>Birth</b>	Null		1	
	not null		8699	
<b>Death</b>	Null		3	0.1
	not null		4269	99.9

### Address 2

Full address 2.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Null		5226	2
	not null		8526	98
<b>Death</b>	Null		2	0.1
	not null		4270	99.9
	Dashed line		1	

### Address 3

Full address 3.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Null		931	10.7
	not null		7769	89.3
<b>Death</b>	Null		0	0
	not null		4272	100

## 4.11 Update and creation dates

### Last Update Date, Last Update User, Last Update User Location

Null in all records.

### Creation date/time (date of registration)

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Distinct	CCYYMMDD	8700	100
	Null		0	
<b>Death</b>	Distinct	CCYYMMDD	4272	100
	Null		0	

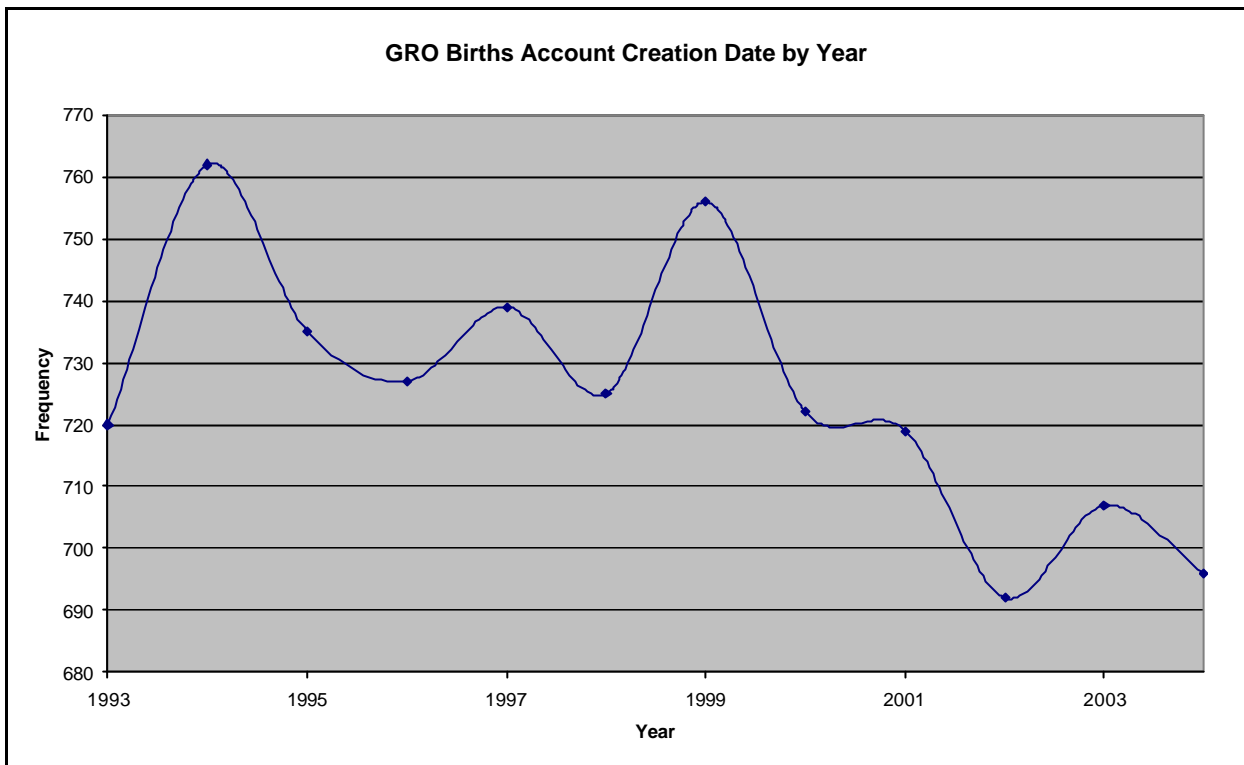


Figure 14 – GRO Births Account Creation Date by Year

4.11.1 For comparison the Births Account Creation Date has also been plotted to the same scale as Date of Birth, Figure 8, and both graphs shown in the following figure. Although the data compares well, there is a surprising difference for year 2004, where c. 80 more accounts have been created than Births recorded. Analysis has shown this is because there is a delay between the birth and the registration, the missing 80 birth records are likely to be registered in 2005. A similar effect is visible for 1992, where c. 80 birth records are shown, but these were not registered until 1993.

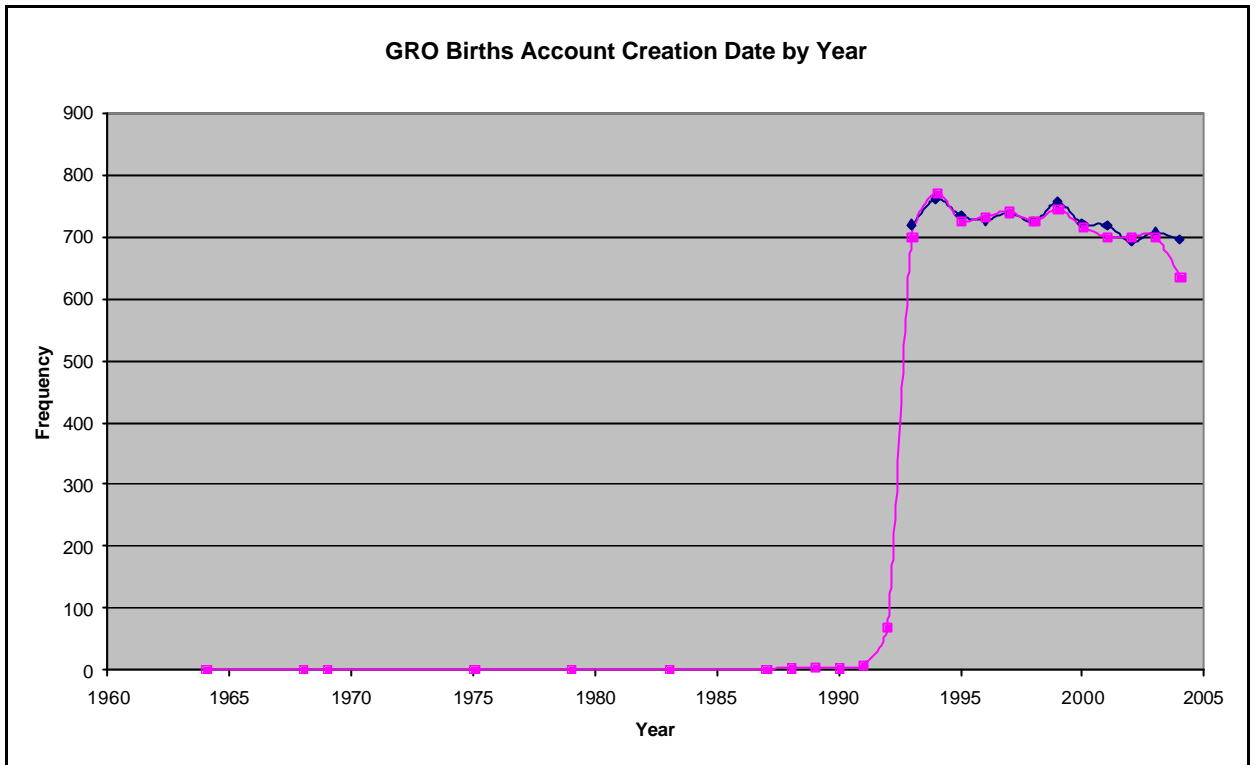


Figure 15 – GRO Births Account Creation Date versus Date of Birth

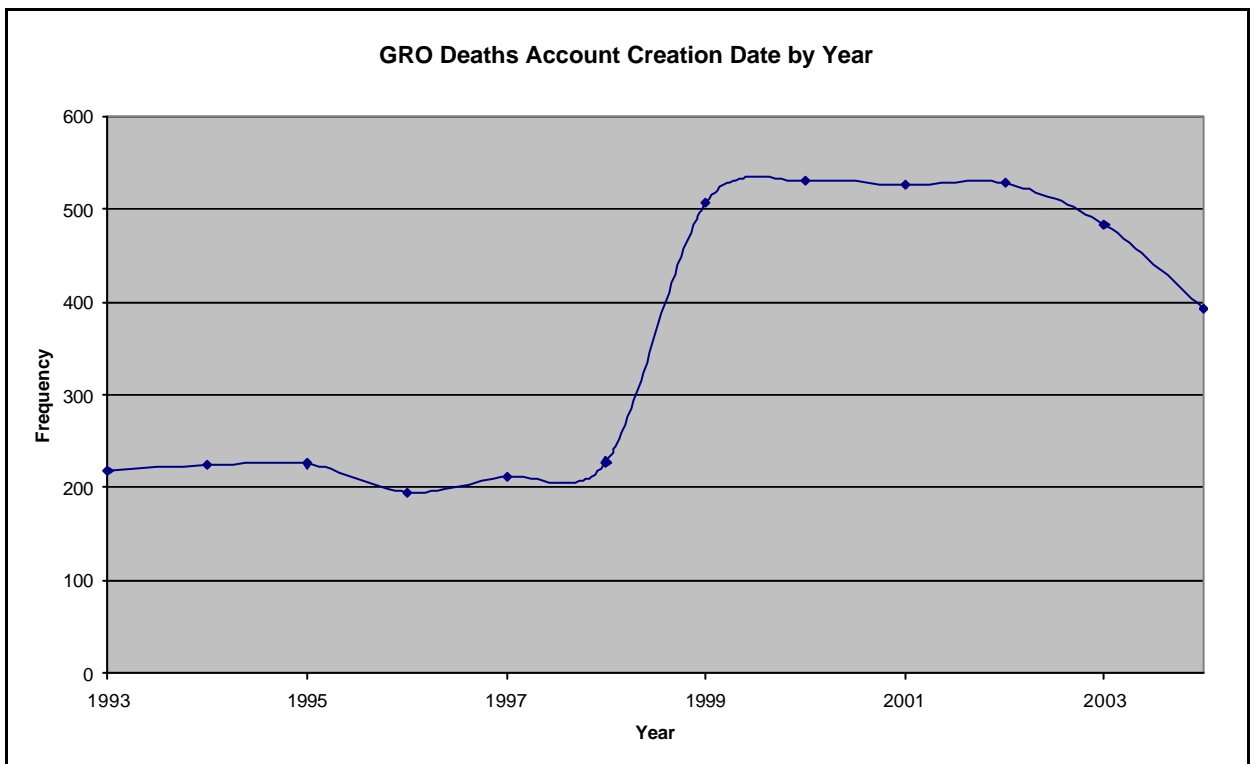


Figure 16 – GRO Deaths Account Creation Date by Year

## 4.12 Miscellaneous data items

### Mainframe Id, M204 userid, Level of Verification, Quality Index,

Null in all records.

### Last Contact Date, Last Contact Time

Null in all records.

### Caution Mark

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Null		8412	96.7
	1		288	3.3
<b>Death</b>	Null		4185	0.1
	1		87	99.9

### Correction Applied Indicator

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Null		8367	96.2
	1		333	3.8
<b>Death</b>	Null		4253	99.6
	1		19	0.4

### Duplicate Record Id

Null in all records.

### Error of Fact or Substance Indicator

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Null		8698	100
	1		2	0
<b>Death</b>	Null		4272	100
	1		0	0

### Ever Manually Examined

Null in all records.

### Missing History Indicator

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	Null		8675	99.7
	1		25	0.3
<b>Death</b>	Null		4269	99.9
	1		3	0.1

### Place of Birth Qualifier

Always null in death records.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8695</b>	<b>99.9</b>
	<b>1</b>		<b>5</b>	<b>0.1</b>
<b>Death</b>	<b>Null</b>		<b>4272</b>	<b>100</b>
	<b>1</b>		<b>0</b>	<b>0</b>

### Place of Birth Qualifier Non Standard Text

Null in all records.

### Registration Corrected Indicator

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8675</b>	<b>99.7</b>
	<b>1</b>		<b>25</b>	<b>0.3</b>
<b>Death</b>	<b>Null</b>		<b>4257</b>	<b>99.6</b>
	<b>1</b>		<b>15</b>	<b>0.4</b>

### Source System (RSS2000, disc or GRONET)

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>4893</b>	<b>56.3</b>
	<b>1</b>		<b>3432</b>	<b>39.4</b>
	<b>2</b>		<b>375</b>	<b>4.3</b>
<b>Death</b>	<b>Null</b>		<b>1654</b>	<b>38.7</b>
	<b>1</b>		<b>2406</b>	<b>56.3</b>
	<b>2</b>		<b>212</b>	<b>5</b>

### Suspense Indicator

Null in all records

### Validation Status

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8673</b>	<b>99.7</b>
	<b>1</b>		<b>27</b>	<b>0.3</b>
<b>Death</b>	<b>Null</b>		<b>4270</b>	<b>100</b>
	<b>1</b>		<b>2</b>	<b>0</b>

### Validation Report Indicator (Disk and BAD System errors)

Null in all records

### Validation Report Indicator (BAD System miskeys only)

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8685</b>	<b>99.8</b>
	<b>2</b>		<b>15</b>	<b>0.2</b>
<b>Death</b>	<b>Null</b>		<b>4271</b>	<b>100</b>
	<b>2</b>		<b>1</b>	<b>0</b>

### Validation Error Code(s)

Can hold multiple error values.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Null</b>		<b>8673</b>	<b>99.7</b>
	<b>Not null</b>		<b>27</b>	<b>0.3</b>
	9A		10	
	NHS		11	
	Other values		6	
<b>Death</b>	<b>Null</b>		<b>4270</b>	<b>100</b>
	8B		1	
	11A, 11B		1	

## 5. Identity duplication

### 5.1 Date of birth, name and address matching

5.1.1 Comparing name, address and date of birth, a number of people were found who appear to have two separate Birth records. 57 such records were found. Extrapolating to the full dataset, we would expect 0.66 +/- 0.17% of records to be duplicates. In a database of 10 million Births, this would equate to between 49,000 and 83,000 duplicates. No people with three Birth records were found.

5.1.2 In summary, therefore, there are 57 (64 + 5 - 12) duplicated GRO records. These appear to be genuine cases of an individual holding two Birth certificates. The total number of GRO person records is 8,697 so the underlying duplication rate is 57 people out of (8,697 - 57) people, or 0.66%.

5.1.3 Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of

$$0.66\% \pm 2sd =$$

$$0.66\% \pm 0.17\%$$

5.1.4 In a database of 10 million people, this equates to between 49,000 and 83,000 people with duplicate (i.e. two) Birth records.

## 5.2 Date of birth and name matching

5.2.1 The number of matches increased to 91 records. Of these:

- 17 records are matches Birth-Death,
- 74 are duplicate births

5.2.2 These included the 5 records that were removed post QAS processing. The rest have very similar addresses and they were not originally identified as duplicates, because QAS could not parse and verify the address. The total number of GRO person records is 8,697 so the duplication rate is taken to be 74 people out of (8,697 – 57) people, or 0.86%.

5.2.3 Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of

$$0.86\% \pm 2sd =$$

$$0.86\% \pm 0.20\%$$

5.2.4 In a database of 10 million people, this equates to between 66,000 and 106,000 people with duplicate (i.e. two) Birth records.

## Appendix C: Data assessment - GRO(S)

---

## 1. Coverage

- 1.1.1 The GROS data covers a period from 1974. Assuming an average age of 74 the expected coverage (for Scotland) is  $30/74$ ths = 41% of the population. For births, this covers the age range 0 to 30 years.
- 1.1.2 The high coverage for s5 (Scotland) is consistent with the 41% expected coverage from the GROS dataset, for both Births and Deaths.

## 2. Field analysis

- 2.1.1 The data comprised the following key fields:

- Unique reference number
- Date of Birth
- Date of Death
- Gender
- Marital Status
- Separate Place of Birth address fields
- Surname and Forename fields
- Separated Address fields
- Date of Registration

- 2.1.2 There were also the following fields:

- Second name (2% populated for Deaths)
- Second address (3% populated for Deaths)

- 2.1.3 As with the other datasets, overall quality is high. Of the key records listed above, nearly 100% of records are populated with the following exceptions:

Place of Birth, 80% of records are null

- 2.1.4 Address quality will be assessed in Lot 2.

- 2.1.5 Analysis of Date of Birth information shows that there does not appear to be a significant use of 1<sup>st</sup> January as a default birth date, either for Births or Deaths.

## 3. Identity duplication

- 3.1.1 Comparing name, address and date of birth, there are no duplicate Birth or Death records. 27 death records were however matched to corresponding Birth records.

## 4. Demographic analysis

- 4.1.1 The GROS sample is split into births and death and contains data for the demographics s1, s2, s5 and s9. The demographics s3, s4, s6 and s7 all fell outside of Scotland and the s8 criteria pre-dated GROS data.
- 4.1.2 Demographic differences derived by Individual Column Analysis were not particularly expected, and none were found.
- 4.1.3 For Gender overall, the Birth data contains 51.7% males, which extrapolates to the full dataset as 51.7% +/- 2%, consistent with a hypothetical 50:50 split between males and females. Death data is also consistent with 50%.

## 5. Data structure

- 5.1.1 GROS data was supplied as eight files, split by births and deaths across the four demographics s1, s2, s5 and s9. s8, which covers data relating to the early 70's, pre-dates the GROS database.
- 5.1.2 The file format was csv but there were no quotes around data values. However, there did not appear to be any occurrences of embedded commas in the data.
- 5.1.3 The data contained the following information per record:
- Unique reference number
  - Date of Birth
  - Date of Death
  - Gender
  - Marital Status
  - Separate Place of Birth address fields
  - Surname and Forename fields
  - Separated Address fields
  - Date of Registration
- 5.1.4 There were also the following fields:
- Second name (2% populated for Deaths)
  - Second address (3% populated for Deaths)

## 6. Statistical Summary

Birth data contains 2733 records and death data contains 2456 records.

### **s1 – Typical Dataset by name**

Surname based sample.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>92</b>	<b>3.4</b>
<b>Deaths</b>	<b>93</b>	<b>3.8</b>

### s2 – Typical Dataset by name

Surname based sample.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>113</b>	<b>4.1</b>
<b>Deaths</b>	<b>77</b>	<b>3.1</b>

### s5 – Covers a rural area in Scotland (postcode)

Scottish postcode area sample of format AAN N\*\*.

Persons with postcode only in address 2 are from an Air Force base.

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>2344</b>	<b>85.8</b>
No qualifying postcode	1	
Postcode in Address 2	14	
<b>Deaths</b>	<b>2274</b>	<b>92.6</b>
Null postcode	4	
Postcode outside criteria	3	

### s9 – Covers issues around nominated date of birth being 1<sup>st</sup> January

Demographic sample for people with a birth date such as 01/01/1976 with the format CCYYMMDD (19760101).

	Frequency	% s0_B or s0_D
<b>Births</b>	<b>184</b>	<b>6.7</b>
<b>Deaths</b>	<b>12</b>	<b>0.5</b>

## 7. Individual Column Analysis

### 7.1 Unique ID

Unique identifier

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>distinct</b>	<b>ANNNNNNNNN N</b>	<b>2731</b>	<b>100</b>
	<b>null</b>		<b>0</b>	
<b>Death</b>	<b>distinct</b>	<b>ANNNNNNNNN N</b>	<b>2456</b>	<b>100</b>

	null		0	
--	------	--	---	--

This not null unique identifier shows 2 people intersecting two demographics.

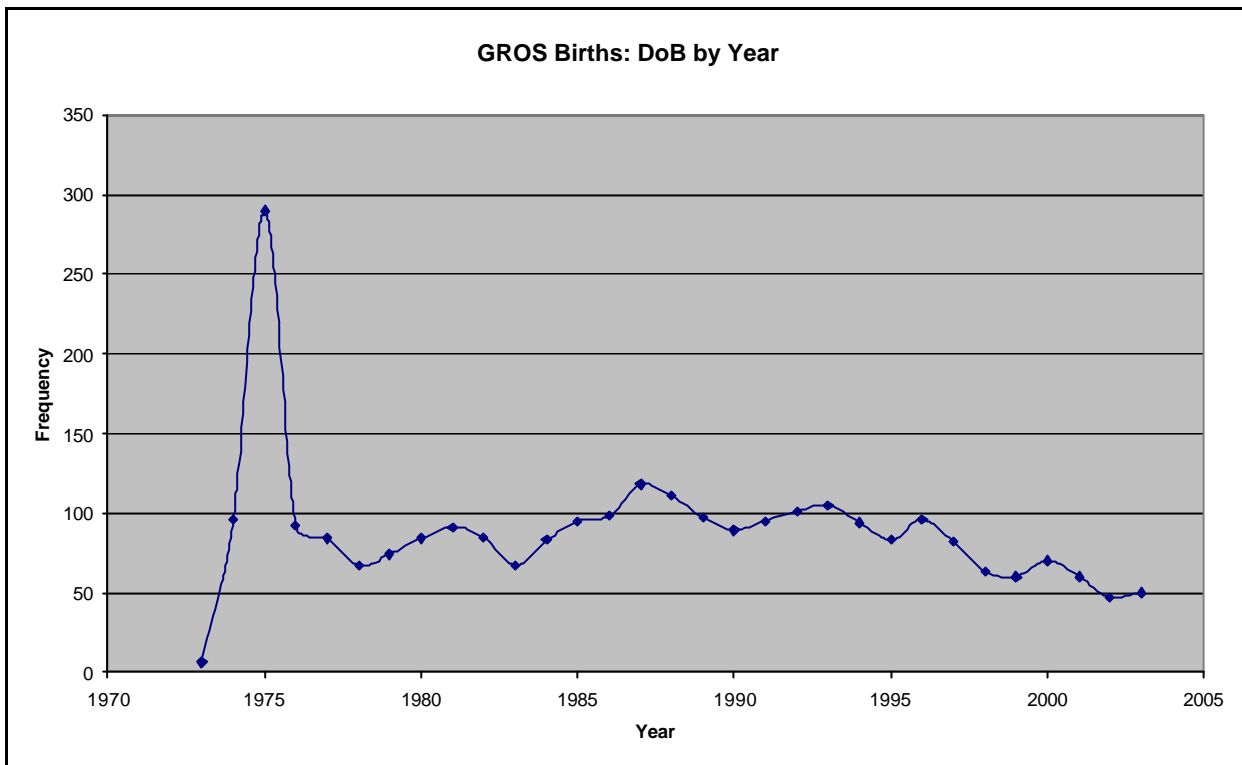
Intersected Demographic	Frequency
<b>Birth</b>	
s1-s5	1
s1-s9	1

**Temp Ref - Null in all records.**

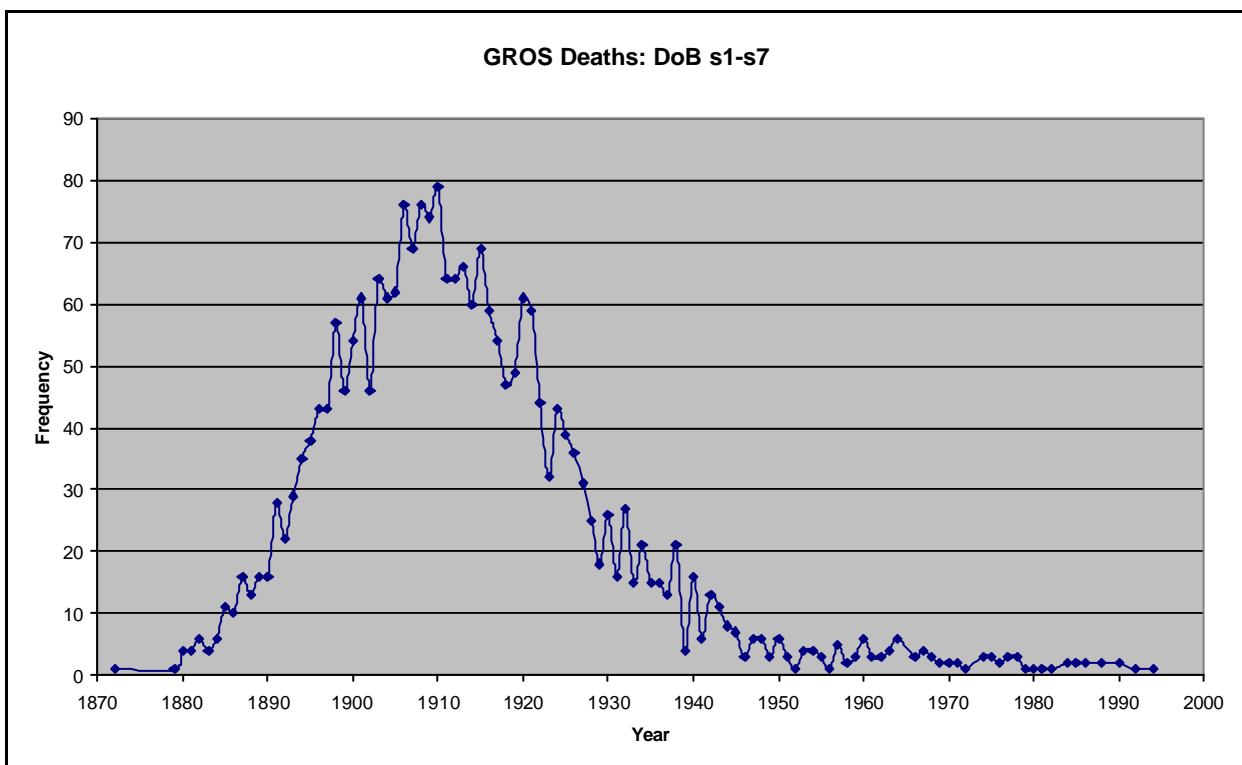
## 7.2 Date of birth

	Value	Format	Frequency	% s0 B or s0 D
<b>Birth</b>	distinct	CCYYMMDD	2733	100
	null		0	
<b>Death</b>	distinct	CCYYMMDD	2456	100
	null		0	

- 7.2.1 The range of Date of Births for GROS Births is from 1973 to 2003 (interestingly no 2004 data was supplied). The range of dates is shown in the following graph. As expected, the shape of the graph closely matches that for Date of Registration. Note that Date of Birth precedes Date of Registration by a few days, and this explains why Date of Birth ranges from 1973, but Date of Registration ranges from 1974. There are 6 records with Date of Birth in 1973: these are all late December births and were registered in early January 1974.
- 7.2.2 The obvious peak in 1975 is caused by the demographic s9 being included in the plot, and is therefore an artefact of the sample data.



7.2.3 The range of Date of Birth for GRO Deaths is from 1872 (1 record) through to 1994. The implication is that no deaths for young people have been recorded in the last 10 years. The s9 DoB demographic has been removed from the graph for clarity.



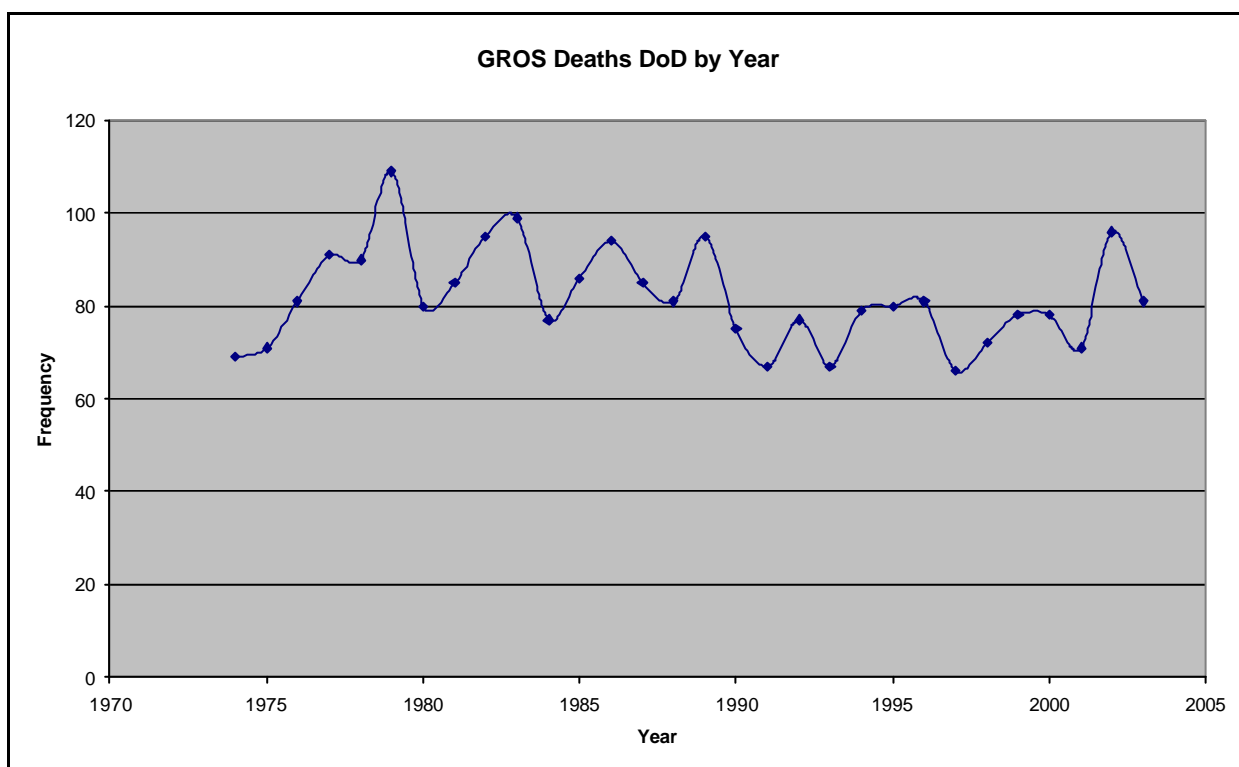
### 7.3 Verified date of birth

Null in all records.

### 7.4 Date of death

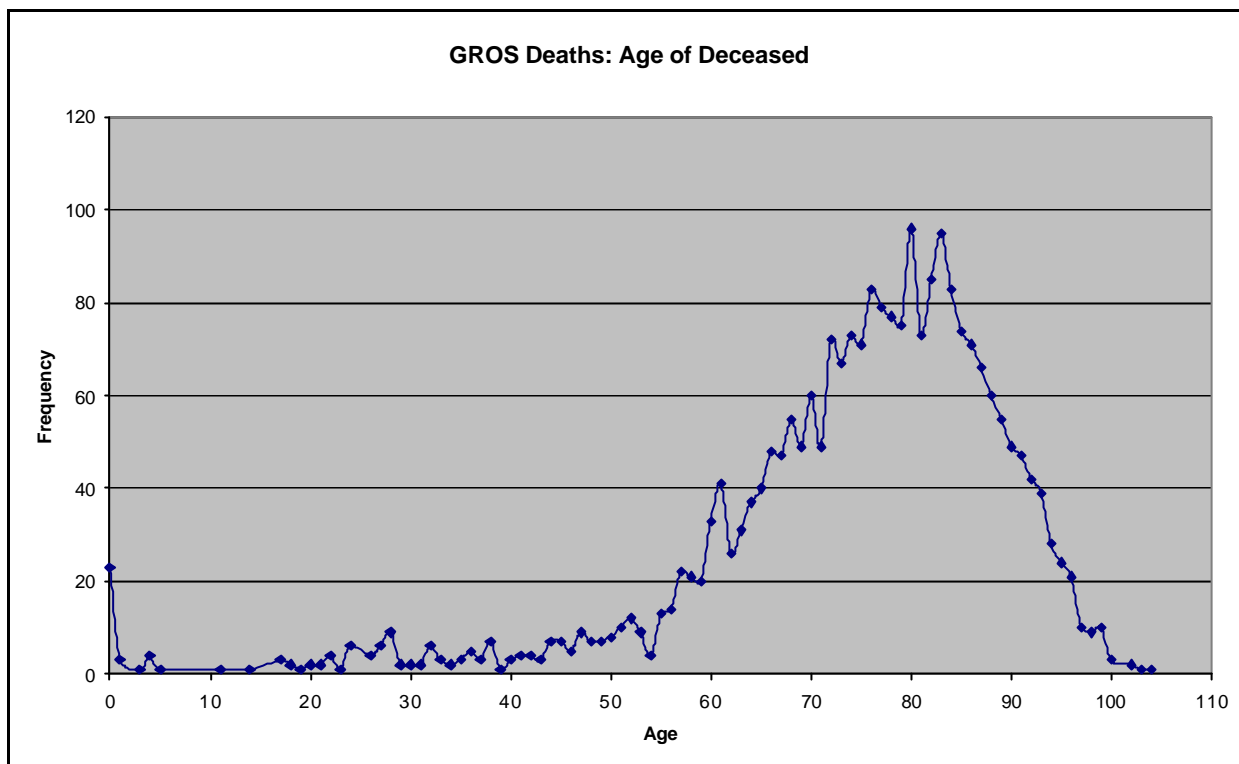
	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>distinct</b>	<b>CCYYMMDD</b>	<b>0</b>	<b>100</b>
	<b>null</b>		<b>2733</b>	
<b>Death</b>	<b>distinct</b>	<b>CCYYMMDD</b>	<b>2456</b>	<b>100</b>
	<b>null</b>		<b>0</b>	

7.4.1 Date of death ranges from 1974 to 2003, as shown in the following graph. No data for 2004 appears to have been supplied. The frequency appears to vary randomly from year to year, which is to be expected.



7.4.2 By subtracting the Date of Birth from the Date of Death, the Age of the deceased can be derived. The frequency of age is plotted in the following graph. As for GRO Deaths, the data is weighted towards the older age group, with an average of 74.0, and there is a notable peak in death rate for infants. There are 23 deaths recorded for infants under 1, of which 91% are for infants under 3 months. Four records indicate a death on the birth day, three of which fall within the 1<sup>st</sup> Jan s9 demographic. Unlike GRO, there is no notable dip around 1917/18.

7.4.3 As indicated under Date of Birth, there are very few deaths recorded for children. Excluding infants, only a further 9 deaths are recorded for under 10's, and none of these occurred since 1995.



### Marital Status

Single for all births.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Single</b>	<b>S</b>	<b>2733</b>	<b>100</b>
	<b>Null</b>		<b>0</b>	<b>0</b>
<b>Death</b>	<b>Divorced</b>	<b>D</b>	<b>67</b>	<b>2.8</b>
	<b>Married</b>	<b>M</b>	<b>1009</b>	<b>41.1</b>
	<b>Single</b>	<b>S</b>	<b>409</b>	<b>16.6</b>
	<b>Widowed</b>	<b>W</b>	<b>968</b>	<b>39.4</b>
	<b>Null</b>		<b>3</b>	<b>0.1</b>

### 7.5 Gender

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>Female</b>	<b>F</b>	<b>1320</b>	<b>48.3</b>
	<b>Male</b>	<b>M</b>	<b>1413</b>	<b>51.7</b>
	<b>null</b>		<b>0</b>	<b>0</b>
<b>Death</b>	<b>Female</b>	<b>F</b>	<b>1233</b>	<b>50.2</b>
	<b>Male</b>	<b>M</b>	<b>1223</b>	<b>49.8</b>
	<b>null</b>		<b>0</b>	<b>0</b>

## 7.6 Place of birth address fields

### Place of Birth Address 1 Place of Birth Address 1

7.6.1 First line of place of birth address often containing being a hospital but also home addresses.

7.6.2 Other values below are other hospitals and home addresses.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>2201</b>	<b>80.5</b>
	Hospital 1		310	11.3
	Hospital 2		88	3.2
	Hospital 3		51	1.9
	Other values		83	3.1
<b>Death</b>	<b>null</b>		<b>2456</b>	<b>100</b>

### Place of Birth Address 2

Second line of place of birth address often of town. Only included in birth data.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>2201</b>	<b>80.5</b>
	Other values		532	19.5
<b>Death</b>	<b>null</b>		<b>2456</b>	<b>100</b>

### Place of Birth Address 3

Third line of place of birth address usually of town where town not in place of birth address 2. Only included in birth data.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>2714</b>	<b>99.3</b>
	Other values		19	0.7
<b>Death</b>	<b>null</b>		<b>2456</b>	<b>100</b>

### Place of Birth Address 4

Fourth line of place of birth address usually of town where town not in place of birth address. Only included in birth data.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	<b>null</b>		<b>2714</b>	<b>99.3</b>
	Other values		19	0.7
<b>Death</b>	<b>null</b>		<b>2456</b>	<b>100</b>

### Place of Birth Postcode

Only included in birth data. Number of null values is the same as place of birth address 1. All meet valid postcode format.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2021	80.5
	Other values		712	19.5
<b>Death</b>	null		2456	100

## 7.7 Name fields

### Name 1 Forename

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		0	0
	not null		2733	100
<b>Death</b>	null		0	0
	not null		2456	100

### Name 1 Surname

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		0	0
	not null		2733	100
<b>Death</b>	null		0	0
	not null		2456	100

### Name 1 Type

Type of record. Child for births. Deceased for deaths.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		0	0
	child		2733	100
<b>Death</b>	null		0	0
	deceased		2456	100

### Name 2 Forename

Alternative forename, but not null values appear to be surnames. Null for births

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2733	100
	Not null		0	0
<b>Death</b>	Null		2451	98.8
	Not null		5	0.2

### Name 2 Surname

Alternative surname. Null for births. The five populated records have the value Former, this would suggest the alternative forename and surname columns actually just hold former surname details.

	Value	Format	Frequency	% s0_B or s0_D
--	-------	--------	-----------	----------------

<b>Birth</b>	null		<b>2733</b>	<b>100</b>
	Not null		<b>0</b>	<b>0</b>
<b>Death</b>	null		<b>2451</b>	<b>98.8</b>
	Not null	Former	<b>5</b>	<b>0.2</b>

## 7.8 Address fields

### Address 1 Line 1

First line off address usually house name, street or in this sample RAF station.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		<b>0</b>	<b>0</b>
	Not null		<b>2733</b>	<b>100</b>
<b>Death</b>	null		<b>0</b>	<b>0</b>
	Not null		<b>2456</b>	<b>100</b>

### Address 1 Line 2

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		<b>7</b>	<b>0.2</b>
	Not null		<b>2726</b>	<b>99.8</b>
<b>Death</b>	null		<b>13</b>	<b>0.5</b>
	Not null		<b>2443</b>	<b>99.5</b>

### Address 1 Line 3

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		<b>1947</b>	<b>71.2</b>
	Not null		<b>786</b>	<b>28.8</b>
<b>Death</b>	null		<b>2051</b>	<b>80.3</b>
	Not null		<b>405</b>	<b>19.7</b>

### Address 1 Line 4

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		<b>2733</b>	<b>100</b>
	Not null		<b>0</b>	<b>0</b>
<b>Death</b>	null		<b>2426</b>	<b>98.8</b>
	Not null		<b>30</b>	<b>1.2</b>

### Address 1 Postcode

Postcode for address 1. All meet valid postcode format.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		<b>19</b>	<b>0.7</b>
	Not null		<b>2714</b>	<b>99.3</b>
<b>Death</b>	null		<b>8</b>	<b>0.3</b>

	Not null		2448	99.7
--	----------	--	------	------

### Address 2 Line 1

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2710	99.2
	Not null		23	0.8
<b>Death</b>	null		2391	97.4
	Not null		65	2.6

### Address 2 Line 2

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2711	99.2
	Not null		22	0.8
<b>Death</b>	null		2391	97.4
	Not null		65	2.6

### Address 2 Line 3

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2728	99.8
	Not null		5	0.2
<b>Death</b>	null		2434	0.9
	Not null		22	99.1

### Address 2 Line 4

Null in all data-set records.

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2733	100
	Not null		0	0
<b>Death</b>	null		2466	100
	Not null		0	0

### Address 2 Postcode

All meet valid postcode format.

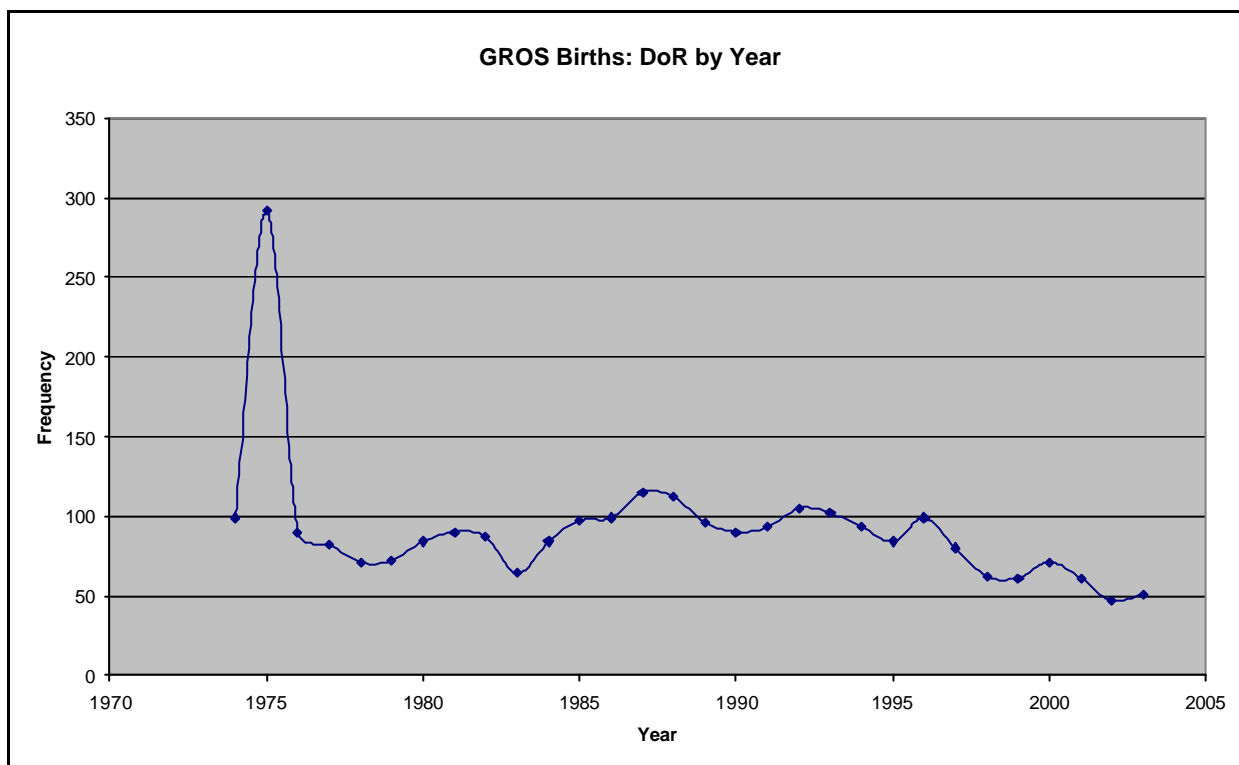
	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		2238	81.9
	Not null		495	18.1
<b>Death</b>	null		2403	97.8
	Not null		53	2.2

## 7.9 Date of registration

Equivalent to creation date

	Value	Format	Frequency	% s0_B or s0_D
<b>Birth</b>	null		0	0
	distinct	CCYYMMDD	2733	100
<b>Death</b>	null		0	0
	distinct	CCYYMMDD	2456	100

7.9.1 Birth registration dates range from 1974 to 2003 (interestingly no 2004 data was supplied). The range of dates is shown in the following graph. As expected, the shape of the graph closely matches that for Date of Birth, indicating that births are registered quickly. The peak in 1975 is caused by the s9 demographic.



## 8. Identity duplication

8.1.1 The identity duplication results are shown in section **Error! Reference source not found., Error! Reference source not found..** A description of the GROS result is given here.

### 8.2 Date of birth, name and address matching

8.2.1 The headline match rate for GROS was 0.47% or 27 records out of the 5,754 split records.

8.2.2 All 27 Review matches are matches between a Birth record and a corresponding Death record. The net match rate is therefore zero. The same information

applies to the PAF result – in other words, no change to the match rate occurred after address cleansing.

### **8.3 Date of birth and name matching**

- 8.3.1 The number of matches increased to 32 records. These all appear to be matches between birth and death records. Inspection of the data indicates these are all genuine matches. The improved match rate is due to poorly formed addresses in the previous matching process failing to match.

## Appendix D: Data assessment - HMRC

---

## 1. HMRC data structure

- 1.1.1 The HMRC data was delivered in three data-files, one containing surname criteria data for s1 and s2, one postcode with all postcode criteria data for demographics s3-s7 and the other based on date of birth criteria s8-9.
- 1.1.2 HMRC provided more documentation than the other stakeholders, including a detailed specification of the extract process and table definitions of the relevant source tables. These proved most helpful in understanding the data format. The data was found to conform to the specification – a fact which was not true with some of the other stakeholders.
- 1.1.3 The file format was tab-delimited. Note that text items were not quoted. There did not appear to be any embedded tables in the textual data however.
- 1.1.4 The data arrived without a demographic id and with date columns in a DD/MM/YYYY format. A data management task was created and run to evaluate and insert a demographic id and correct date formatting to CCYYMMDD.
- 1.1.5 The data comprised the following key fields:
- Unique reference number (NINO)
  - Date of Birth
  - Date of Death
  - Gender
  - Date of Entry (around the date on which person was 16)
  - Date of Registration (date person registered on NIRS)
  - Name with separate Surname and Forename fields
  - Name Last Update Date
  - Address with separated Address fields
  - Address Last Update Date
- 1.1.6 The HMRC data structure was more complex than the other stakeholders. HMRC data, unlike DVLA, UKPS, GRO and GROS, holds its data in a normalised way, i.e. with a historical list of names and addresses for each person record. The data was extracted with one name and address pair per line, multiple lines per person. The following observations were made:
- The total number of records for each NINO is the Cartesian product of the number of name records and the number of address records extracted. Hence if 2 name records and 3 address records were extracted, there will be 6 records for that NINO in the file.

The extract format significantly complicates the concept of a “record”, making it much harder to derive statistics on a per record (i.e. NINO) basis, especially for name and address data. The later is more fully explored under ‘HMRC record currency analysis’.

- HMRC record currency analysis.
- The demographic id had been omitted so had to be inferred.
- The extract by demographic was done differently than for other stakeholders, resulting in only partial data for each NINO record. In other words, some names and addresses that existed for the NINOs extracted were missing. This arose for the name and address demographics, where the algorithm used was as follows, illustrated for the name demographic (address in brackets):
  - Select all names (addresses) that match the demographic (i.e. ignore names (addresses) that don't match)
  - For each record, extract NINO data and add to each name (address) record
  - For each NINO, extract address (name) data and create a name-address record for each address (name) found.
  - In other words, names (addresses) that existed for each selected NINO were omitted.
- Each name and address date range is independent. HMRC has matched every name with every address, rather than attempt to create name-address pairs that overlapped in time. This was not unreasonable, as the date stamps only refer to the date when HMRC recorded a change in name or address, not necessarily the date from which a particular name or address was used. Hence, by combining all possible name-address pairs as HMRC has done will maximise the chance of matching records with other datasets, even though it generates additional name-address pairs that were never used in reality. Note that we have followed a similar approach in combining multiple-name and address pairs from the other stakeholders.

## 2. Statistical summary

- 2.1.1 The data contains multiple records per person. Each address and name combination creates a new record.
- 2.1.2 The total number of records in the HMRC data including multiple records per person is 208025. These multiple records skew comparisons with other data sources. To correct this, a data management job was run (IR\_RemoveDuplicateNINO\_s0.mpj) on s0 to produce a data file containing one record per each combination of demographic / NINO with no names or addresses. The total number of distinct records is 93612.

2.1.3 In the next section:

- Tabular data is as produced by ISI Profiler on either the 208025 or the 93612 (Distinct) record files as indicated.
- Graphical data is produced by Excel on an extract of 65535 records. For DoB, DoD, DoE and DoR, these are extracted from the Distinct file, and represent 100% of s1 and s2 (10118 records) and 73% of s3-s7 (55418/75875). There are no s8 or s9 records in this extract.

	Demographic	Frequency	%	Distinct NINO	%
s1	Typical Dataset by name	13588	6.5	5841	6.2
s2	Typical Dataset by name	10332	4.7	4277	4.6
s3	Typical suburban dataset by geographic area (postcode and area name)	10332	4.7	4277	4.6
s4	Covers name issues and address issues on houses that have been converted into flats. (postcode)	48681	23.4	28517	30.5
s5	Covers a rural area in Scotland (postcode)	18487	8.8	8062	8.6
s6	Covers issues around Welsh names and addresses (postcode and area name)	11113	5.3	5756	6.2
s7	Covers issues related to high density urban areas and high rise flat blocks	33026	18.8	19086	20.4
s8	Dataset by specific date of birth	13041	6.3	2922	3.1
s9	Covers issues around nominated date of birth being 1 <sup>st</sup> January	22956	11	4694	5.0

### 3. Individual column analysis

#### 3.1 National Insurance Number (Nino)

3.1.1 National Insurance Number is unique per person

3.1.2 The data contains multiple records per person.

3.1.3 From the distinct HMRC data the following 29 intersections of records over demographics was found.

Intersected Demographic	Frequency
s1-s7	1
s1-s9	1
s2-s4	3
s2-s7	1
s2-s8	1
s3-s7	1
s3-s9	1
s4-s5	1
s4-s7	2
s4-s8	1
s4-s9	10
s5-s8	1
s6-s9	1
s7-s9	4
<b>Total</b>	<b>29</b>

Value	Format	Frequency	Distinct	Distinct % s0
<b>Null</b>		<b>0</b>	<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>93583</b>	<b>100</b>

## 3.2 Nino suffix

3.2.1 Each Nino has one distinct suffix.

3.2.2 The null records all have an account status of redundant.

Value	Format	Frequency	Distinct	Distinct % s0
<b>null</b>		<b>3</b>	<b>3</b>	<b>0</b>
<b>not null</b>		<b>208022</b>	<b>93609</b>	<b>100</b>
A		52540	23803	25.4
B		50797	23007	24.6
C		52535	23319	24.9
D		52150	23480	25.1

## 3.3 Account status

3.3.1 The Values shown in the following table are those specified by HMRC with no further explanation.

Value	Format	Frequency	Distinct	Distinct % s0
<b>null</b>		<b>0</b>	<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>93612</b>	<b>100</b>
Full Live	0	207059	92829	99.2
Pseudo lom Post	1	0	0	0
Full lom Post 86	2	17	13	0
Full Cancelled	3	3	3	0

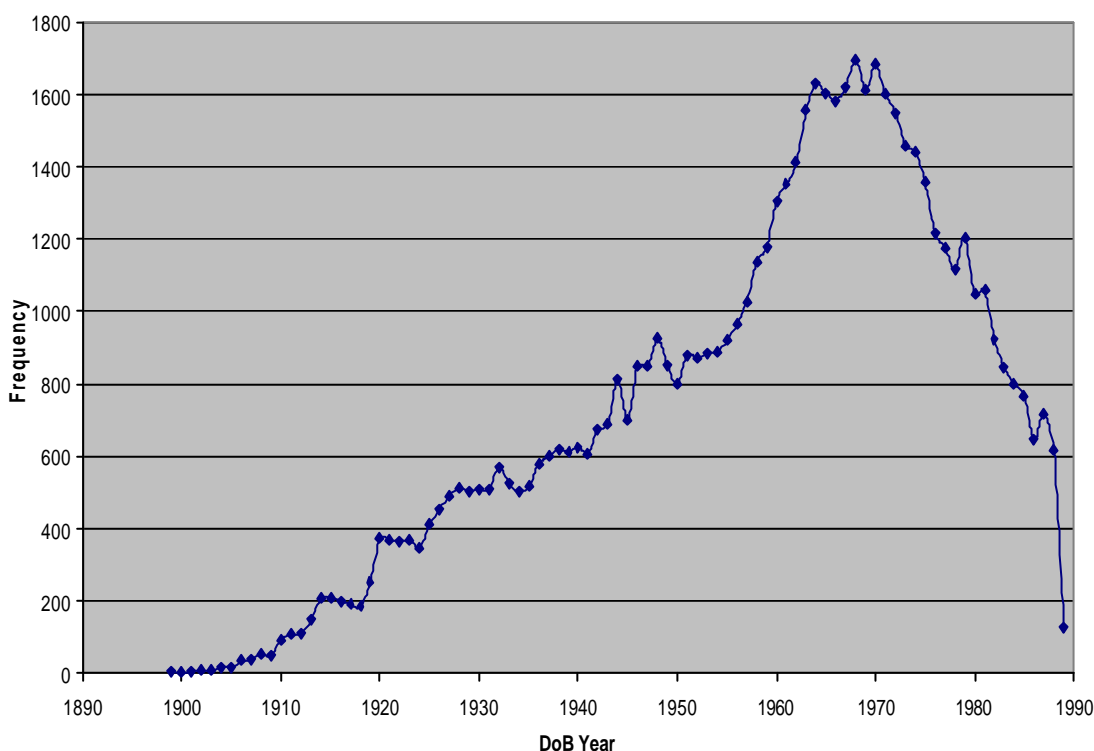
Full Amalgamated	4	86	72	0.1
Full Administrative	5	0	0	0
Pseudo Weeded	6	0	0	0
Pseudo Amalgamated	7	0	0	0
Pseudo Other	8	0	0	0
Redundant	9	860	695	0.7
Conversion Rejection	10	0	0	0

### 3.4 Date of birth

3.4.1 Before analysis the date of birth was reformatted to CCYYMMDD.

Value	Format	Frequency	Distinct	Distinct % s0
Null	CCYYMMDD	25	24	0
not null		208000	93588	100

3.4.2 Figure 17 shows a frequency distribution of the first 65536 records (the limit of MS Excel) from the “Distincts” file of 93612 records.

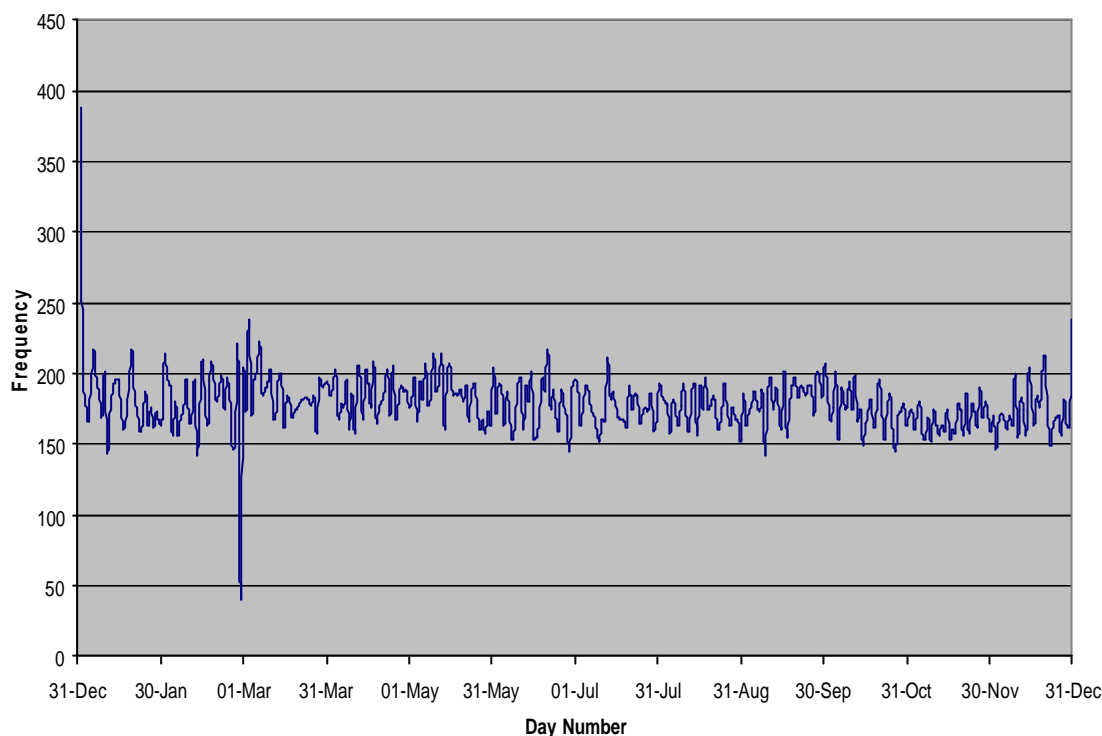


**Figure 17 - HMRC DoB by Year (65536)**

3.4.3 Note that the DoB data has an upper cut-off at 1988. This is to be expected, as the HMRC database only included people once they are 16 years old (or up to three months before), and the data extract was made at the end of 2004.

3.4.4 The same data (i.e. s1-s7) is also be plotted to show frequency of particular days in the year, to indicate if there is a particular peak at 1<sup>st</sup> Jan in any year (Figure

18). The year data was removed, the remaining month-day information sorted, and the occurrence of each of the 366 values counted and plotted in the graph below. (To simplify the process, the month-day data was combined with the year value 2004 (a leap year) to assist the production of the x-axis.)



**Figure 18 - HMRC DoB by Day Number (65536)**

3.4.5 Note the peak at 1<sup>st</sup> January and dip at 29<sup>th</sup> February. The average frequency for any given date is 179, and the standard deviation (sd) (ignoring data for 1/1 and 29/2) is 16 (8.9%) – nearly all the data falls within 1 sd, and most of it within 2. There is nothing else outside 3 sd. The theoretical average and sd for a probability of 1/365 (0.0027) and sample size as specified are 179 and 13 respectively – very close to the measured values.

3.4.6 The peak at 01/01 is 389, over twice the average value ( $389/179 = 2.17$ ), and some 13 sd larger. This is very unlikely to be random and we assume it is a real feature of the data, suggesting that approximately half of the 01/01 births are default values. The probability of a date being 01/01 is  $389 / 65535 = 0.0059$ . The theoretical sd for this value is 20 (5%), and so in the full population, we might expect an occurrence rate of  $0.0059 \pm 0.0006$  (2sd) (i.e. within 10%). We can also calculate the uncertainty in the ratio (1<sup>st</sup> Jan occurrence over average occurrence) by combining the statistical uncertainties in each term (i.e. combining 8.9% and 5%). This is done using the standard “sum of squares” formula to produce an sd of about 10% (of the value of the ratio) or an error range of  $2.17 \pm 0.44$  (2sd). We do not know the full population size, but we do know the s8 and s9 sizes for the full population, i.e. 2922 and 4694 respectively, and the ratio  $s9/s8$  should be similar to 2.17. In fact, taking into account the variation in birth rate between the two demographics, the ratio is derived as  $2.4/1.3$  or 1.91. The

two values (2.17 and 1.91) are hence well within 1sd of each other, and therefore consistent.

3.4.7 The corollary is that  $1/2.17 = 46\% \pm 9\%$  (2sd) of 1<sup>st</sup> January dates are in fact correct, whereas  $54\% \pm 9\%$  are default dates.

3.4.8 The dip at 29/02 occurs because only 1 year in four is a leap year, and hence we would expect the number of occurrences of 29/02 to be approximately  $\frac{1}{4}$  of the average, i.e. 45. In fact there are 40 occurrences of 29/02, which is slightly lower than expected. However, the theoretical standard deviation for 40 occurrences out of 65,535 records is 6.3, so the measured value 40 is well within 1sd, so is consistent.

### 3.5 Date of birth status

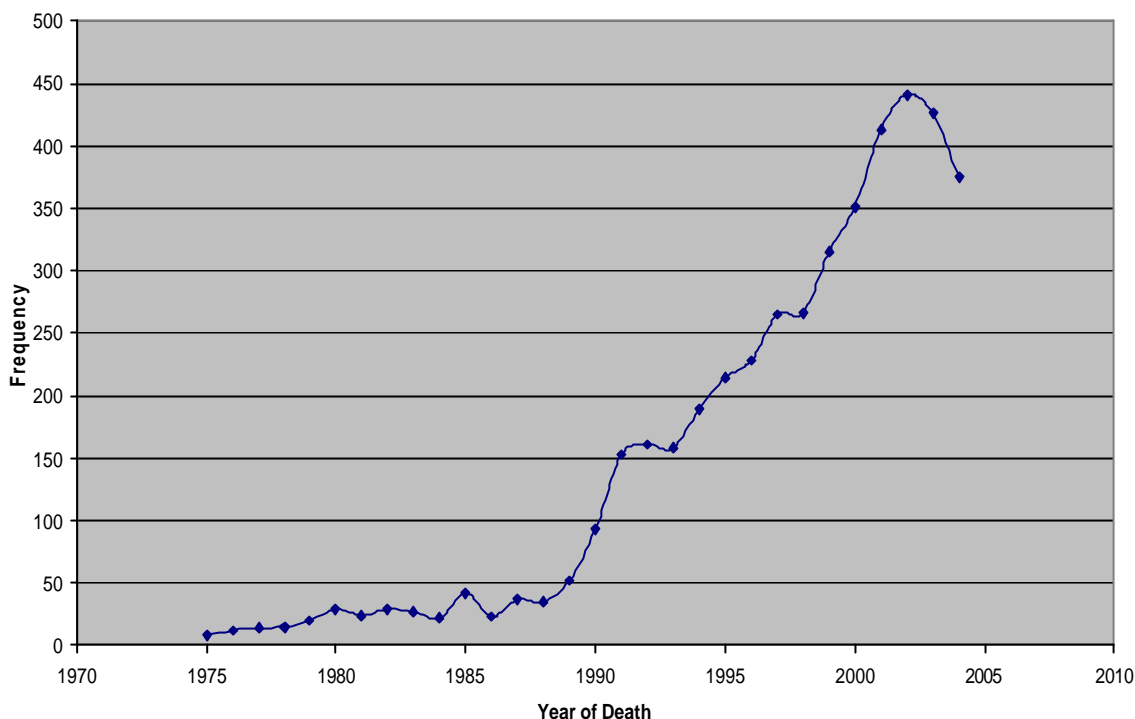
Value	Format	Frequency	Distinct	Distinct % s0
<b>null</b>		<b>0</b>	<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>93612</b>	<b>100</b>
Unverified	0	60302	27952	29.9
Verified	1	147717	65656	70.1
Not Known	2	0	0	0
COEG Confirmed	3	6	4	0

### 3.6 Date of death

3.6.1 Before analysis the date of death was reformatted to CCYYMMDD.

Value	Format	Frequency	Distinct	Distinct % s0
<b>Null</b>	CCYYMMDD	<b>199043</b>	<b>87426</b>	<b>93.4</b>
<b>not null</b>		<b>8982</b>	<b>6186</b>	<b>6.6</b>

3.6.2 Figure 19 shows DoD for the 65536 Distinct extract data. Note that this held 4437 populated DoD records (6.8%): in other words, the figure is a graph of these 4437 records.



**Figure 19 - HMRC DoD by Year (65536)**

3.6.3 The graph shows a high level of recent deaths, peaking in 2000, and a much lower rate of death pre 1990. This potentially represents the fact that the coverage of the HMRC database prior to 1990 was significantly smaller than it is today.

### 3.7 Date of death status

3.7.1 The values 0 and 1 are undefined.

Value	Format	Frequency	Distinct	Distinct % s0
<b>null</b>		<b>25951</b>	<b>12776</b>	<b>13.7</b>
<b>not null</b>		<b>182074</b>	<b>80836</b>	<b>86.3</b>
	0	173399	74862	80
	1	8675	5974	6.3

### 3.8 Gender

3.8.1 The N value is undefined.

Value	Format	Frequency	Distinct	Distinct % s0
<b>Female</b>	<b>F</b>	<b>117566</b>	<b>45568</b>	<b>56.5</b>
<b>Male</b>	<b>M</b>	<b>90443</b>	<b>48029</b>	<b>43.5</b>
	<b>N</b>	<b>9</b>	<b>9</b>	<b>0</b>
<b>null</b>		<b>7</b>	<b>6</b>	<b>0</b>

### 3.9 Date of entry

- 3.9.1 Before analysis the date of entry was reformatted to CCYYMMDD.
- 3.9.2 The date value 17530102 appears to be a default date.
- 3.9.3 The date values 19880115 and 19910101 both appear in over 5% of records.

Value	Format	Frequency	Distinct	Distinct % s0
null	CCYYMMDD	886	711	0.4
not null		207139	92901	99.2
17530102		785	623	0.7
19890115		10875	2268	2.4
19920101		10668	1934	2.1

- 3.9.4 DoE has been plotted by year in Figure 20. This is based on the 65535 extract, of which 65048 records were populated (99.2%). Furthermore, there were 474 occurrences (0.7%) of the 1753 date which were treated as null, taking the total number of valid records down to 64574 (98.5%).

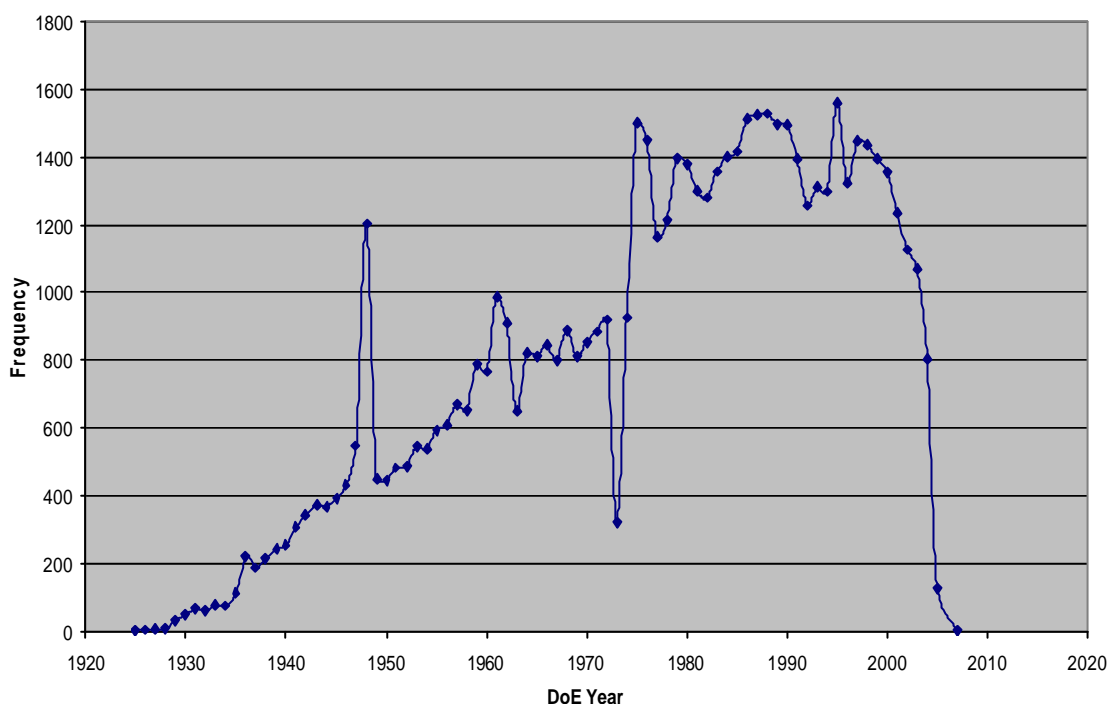


Figure 20 - HMRC DoE by Year (65536)

- 3.9.5 The data ranges from 1925 through to a few values in 2005 (127 records) and even one in 2007 (which may be a typographical error in the data). Apart from a couple of obvious peaks and troughs, the data rises broadly linearly up to a peak in 1990. There is a major peak for 1948 (1204 records), a trough for 1973 (322 records) and peak for 1975 (1501 records). The overall maximum occurs for 1995 with 1560 records. The future dates are tabled below:

Value	Format	Frequency	Percentage Extract
	200501??	20	0.03

	200502??	23	0.04
	200503??	17	0.03
	200704??	1	<0.01

3.9.6 Information about this field from HMRC is as follows. It is meant to be the date at which the individual became liable to UK NI. Usually this would be on or around the individual's 16th birthday for UK residents, or it would be more likely to be Date of Entry into the UK for a foreign national. The date 02/01/1753 is a default given at the move from NIRS1 to NIRS2 if the data field did not hold a relevant value on NIRS1. The 1926 dates usually are on or around the 16th birthday for people who were born in 1910 - even though NI did not theoretically start until 1948. People are registered a few months before their 16th birthday which explains the Date of Entry dates slightly in the future.

3.9.7 This also explains the high frequency of the dates 19890115 and 19920101 which correspond to the s8 and s9 demographics, but 16 years out. Note that these dates

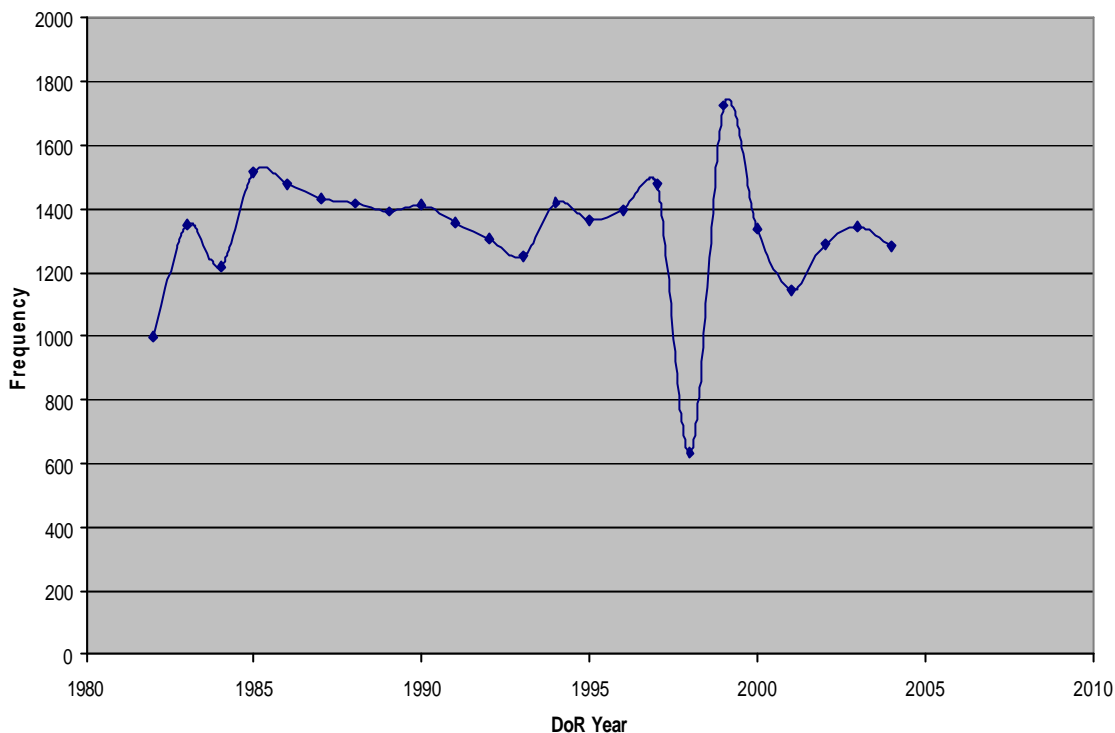
- have been altered to hide the true value for s8 and s9
- and were omitted from the 65536 sample plotted on the graph.

### 3.10 Date of registration

3.10.1 Before analysis the date of registration was reformatted to CCYYMMDD.

Value	Format	Frequency	Distinct	Distinct % s0
null	CCYYMMDD	89428	46577	49.8
not null		118597	47035	50.2

3.10.2 DoR has been plotted by year in Figure 21. This is based on the 65535 extract, of which 30556 records were populated (47%).



**Figure 21 - HMRC DoR by Year (65536)**

3.10.3 The data ranges from 1982 to 2004. Note the dip in 1998 of 633 records, and the peak the following year of 1726 records. Otherwise the graph is moderately constant over a 23 year period with an average value of 1329 records per year (4.3%).

3.10.4 Information from HMRC about this field is as follows. DoR is the date the individuals were recorded on NIRS for the first time. This is usually a few months before the 16th birthday for UK nationals, although this will be different for adult registrations. Before 1982 this information was not recorded so that generally people with a DOB of prior to 1966 will not have this field filled in.

### 3.11 Name sequence number

3.11.1 The order in which a name has been used for a person.

Value	Format	Frequency	Percentage s0
<b>Null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>100</b>
1		146199	70.2
2		39707	19.1
3		12548	6.0
4		5225	2.5
5		2101	1.0
6		965	0.5
7		489	0.3
8		253	.0.1
9		136	0.1
10		106	0.1
11-19		296	0.1

### 3.12 Name elements

#### Name Type

3.12.1 Name type 2 is used for addressing correspondence while name type 1 is used elsewhere. One person can have only one current name type1 and one current name type 2.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>Name 1</b>		<b>203419</b>	<b>97.8</b>
<b>Name 2</b>		<b>4606</b>	<b>2.2</b>

#### Title

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
Not Known	0	32439	15.6
Mr	1	84128	40.4
Mrs	2	46153	22.2
Miss	3	39409	18.9
Ms	4	5458	2.6
Dr	5	388	0.2
Rev	6	50	0.1

#### Forename 1

3.12.2 First forename or initial.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>27542</b>	<b>13.2</b>
<b>not null</b>		<b>118597</b>	<b>86.8</b>
Single initial		1280	0.6

## Forename2

3.12.3 Second forename or initial.

Value	Format	Frequency	Percentage s0
null		27542	13.2
not null		118597	86.8

## Surname

3.12.4 Records with a null surname all had a not null requested name.

Value	Format	Frequency	Percentage s0
null		744	0.4
not null		207281	99.6

## Requested Name

3.12.5 Records contain full name sometimes with title.

Value	Format	Frequency	Percentage s0
null		204578	98.3
not null		3447	1.7

## Name Start Date

3.12.6 There is a large number with a value which appears to be a default date (17530102).

Value	Format	Frequency	Percentage s0
null	CCYYMMDD	0	0
not null		208025	100
17530102		27765	

3.12.7 Name Start Date appears to range from December 1996, although there are a few records from earlier than this, as shown in the following table.

Year	Frequency
1943	1
1953	1
1978	1
1983	2
1987	2
1989	1
1992	6
1994	3
1995	1
199612..	thousands

3.12.8 By inspecting the first name for a person, i.e. with Name Sequence Number = 1, we would expect Date of Registration, Name Start Date and Address Start Date

to be roughly the same. In other words, the first name and the first address are recorded at the time of the registration on the HMRC database. However, we have observed that in most cases, the Address Start Date is approximately the same as the Date of Registration, whereas the Name Start Date is only the same when Date of Registration is 1997 or later. We conclude from these observations, and from the use of the default 17530102 date, that name history information was only recorded in the system from December 1996. The handful of Name Start Date records pre-dating December 1996 were probably amended manually for particular business reasons.

### Name End Date

Value	Format	Frequency	Percentage s0
<b>null</b>	<b>CCYYMMDD</b>	<b>149618</b>	<b>71.9</b>
<b>not null</b>		<b>58407</b>	<b>28.1</b>
17530102		26674	
19980116		1423	

3.12.9 There is a large number with a value which appears to be a default date (17530102) which often coincides with the same value for Name Start Date.

3.12.10 It is assumed that the null end date indicates a current name. However, it is not clear what interpretation should be applied to the default date 17530102. Inspection of a selection of records seems to indicate the following:

- The 17530102 dates appear to exist for records registered prior to 1998
- There seem to be no occurrences for Name Sequence Number = 1. It is always 2 or higher. (Note that there are occurrences of Name Sequence Number = 1 for Name Start Date of 17530102, where Name End Date is null).
- In some cases, Name Sequence Number = 1 has a null End Date, even though there is a second name with Start and End Dates set to 17530102. In other words, the first name is still current. It is as if the second record is informational, and never used.
- In some cases, there is a first and third name (Name Sequence Number = 1 or 3) with valid start and end dates, while the second name has the default 17530102 dates. In this case, the second name is assumed to be not current.

3.12.11 In summary, it would appear to be safer to assume that the default date should not be treated as null – in other words, every record should have a null Name End Date record indicating the current name: it never appears to be the record with End Date set to 17530102.

## 3.13 Address elements

### Address Sequence Number

3.13.1 The order in which an address has been used for a person.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>100</b>
1		82136	39.5
2		48655	23.4
3		27619	13.3
4		15617	7.5
5		9640	4.6
6-68		24358	11.7

### Address Source

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>101258</b>	<b>48.7</b>
<b>not null</b>		<b>106767</b>	<b>51.3</b>
Not Known	0	95331	45.8
Customer	1	3	0
Relative	2	0	0
Employer	3	0	0
HM Revenue and Customs	4	6949	3.3
Other Govt Dept	5	4484	2.2
Other Third Party	6	0	0

### Country Code

3.13.2 Many records have a country code outside of UK but a UK address

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>11</b>	<b>0</b>
<b>not null</b>		<b>208014</b>	<b>100</b>
Not Known/Specified	0	6086	2.9
Great Britain	1	198249	95.3
Eng/Sco/Wal/N Ire/R Ire	114/115/116/8/117	3007	1.4
Abroad Not Known	248	297	0.2
Other Values		386	0.2

### Address Type

3.13.3 Not known refers to the address type being unknown not the address itself

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>100</b>
Not Known	0	33	0
Residential	1	207820	99.9
Correspondence	2	172	0.1

### Address Status

#### 3.13.4 Dead Letter Office (DLO) needs to be explained

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>76984</b>	<b>37.0</b>
<b>not null</b>		<b>131041</b>	<b>63.0</b>
Not DLO	0	122060	58.7
DLO	1	8981	4.3
No Fixed Abode	2	0	0

### Address Line 1

#### 3.13.5 First address line commonly containing house or flat number or name and road name

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>608</b>	<b>0.3</b>
<b>not null</b>		<b>207417</b>	<b>99.7</b>

### Address Line 2

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>605</b>	<b>0.3</b>
<b>not null</b>		<b>207420</b>	<b>99.7</b>

### Address Line 3

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>57277</b>	<b>27.5</b>
<b>not null</b>		<b>150748</b>	<b>72.5</b>

### Address Line 4

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>156886</b>	<b>75.4</b>
<b>not null</b>		<b>51139</b>	<b>24.6</b>

### Postcode

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>3964</b>	<b>1.9</b>
<b>not null</b>		<b>204061</b>	<b>98.1</b>

### Address Start Date

#### 3.13.6 Large number of records with address start date of 19811129.

Value	Format	Frequency	Percentage s0
<b>null</b>	<b>CCYYMMDD</b>	<b>0</b>	<b>0</b>
<b>not null</b>		<b>208025</b>	<b>100</b>
19811129		8958	4.3
20021001		1596	0.8

### Address End Date

3.13.7 A null end date should denote a current address

Value	Format	Frequency	Percentage s0
<b>null</b>	<b>CCYYMMDD</b>	<b>80334</b>	<b>38.6</b>
<b>not null</b>		<b>127691</b>	<b>61.4</b>
20020930		1860	0.9

3.13.8 Record currency analysis, based on Name and Address End Dates is given in the next section.

### Address Confirmed

3.13.9 A null end date should denote a current address

Value	Format	Frequency	Percentage s0
<b>null</b>	<b>CCYYMMDD</b>	<b>40872</b>	<b>19.6</b>
<b>not null</b>		<b>167153</b>	<b>80.4</b>
19811129		7733	3.7
20040929		1503	0.7
20021001		1474	0.7

## 4. HMRC record currency analysis

### 4.1 Scope and analysis of data

4.1.1 As previously explained, the HMRC data contains historical name and address information for each National Insurance Number record. Some analysis of this information has been carried out, although there is more than could still be done as will be explained.

4.1.2 The analysis attempts to answer the following questions:

- What are the numbers of current records by demographic?
- How often do people change address?
- How often do people change name?

### 4.2 Current records by demographic

4.2.1 The HMRC data has multiple name and address data for each NINO.

4.2.2 For the s1 and s2 name demographics, name data has been provided by HMRC only where it matches the extract criteria. Some of these names are current, but some are historic. Not all NINOs extracted will therefore include the current name.

4.2.3 Similarly for the s3-s7 address demographics, address data has been provided by HMRC only where it matches the extract criteria. Some of these addresses are current, but some are historic. Not all NINOs extracted will therefore include the current address.

4.2.4 Only for the s8 and s9 dates demographic are full name and address histories provided.

4.2.5 The counts made on distinct NINO and presented in v1.0 of this document did not take the name or address currency into account. We therefore suppose that the true coverage represented by the record count is greater than if we had only included current records. A current record is defined as one where the Name and Address End Dates are null. Furthermore, Date of Death is also required to be null. The following table presents the number current records compared against the “distinct” record count.

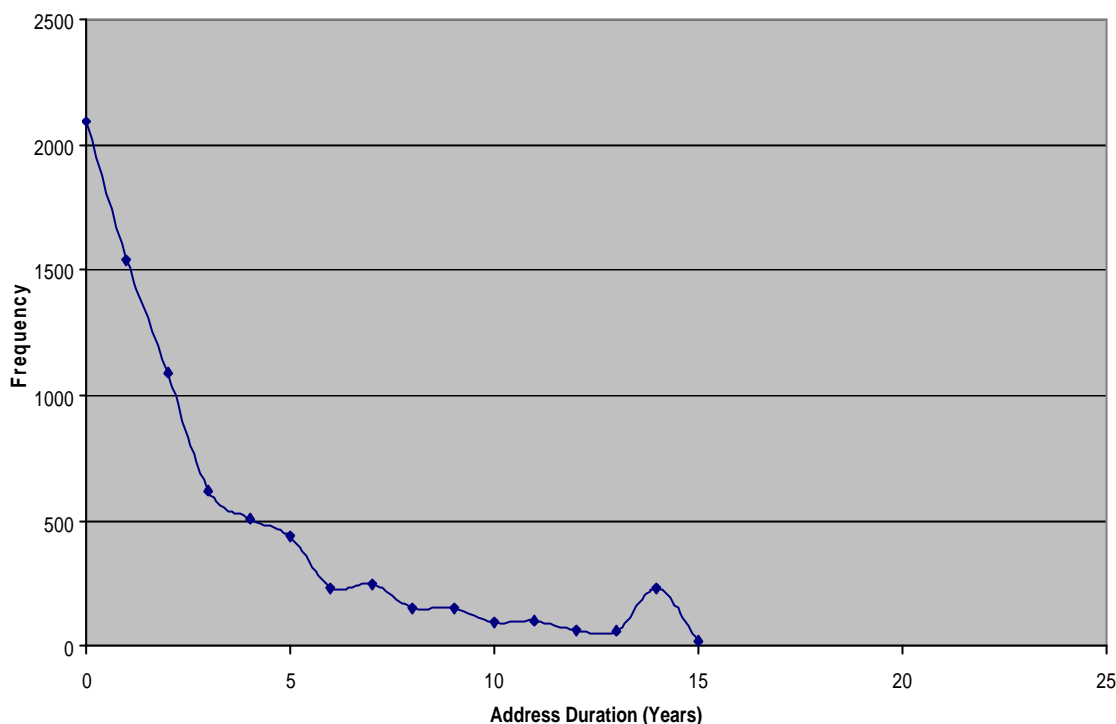
Dem ID	Distinct NINO Count	Number of Current Records	Percentage	Adjusted
S1	5841	4582	88%	4472
S2	4277	3652	87%	3564
S3	14457	8059	62%	7866
S4	28517	15734	59%	15357
S5	8062	4594	64%	4484
S6	5756	3532	69%	3447
S7	19086	9872	59%	9635
S8	2922	2984	102.1%	2922
S9	4694	4825	102.8%	4694
S0	93612	57834		56441

4.2.6 We see that the Name demographic numbers are reduced by about 12-13%, the address demographics are reduced by 30-40%, but surprisingly, the date demographics are increased by 2-3%. This latter result is surprising and implies that 2-3% of NINOs have two current name or address records. A further analysis could be carried out to both investigate these and remove them from the statistics. However, insufficient time has been available for this, and instead, all results have been reduced by 2.5% to account for the assumed duplication. The reduced figures are shown in the final column above labelled “Adjusted”.

### 4.3 Change of address statistics

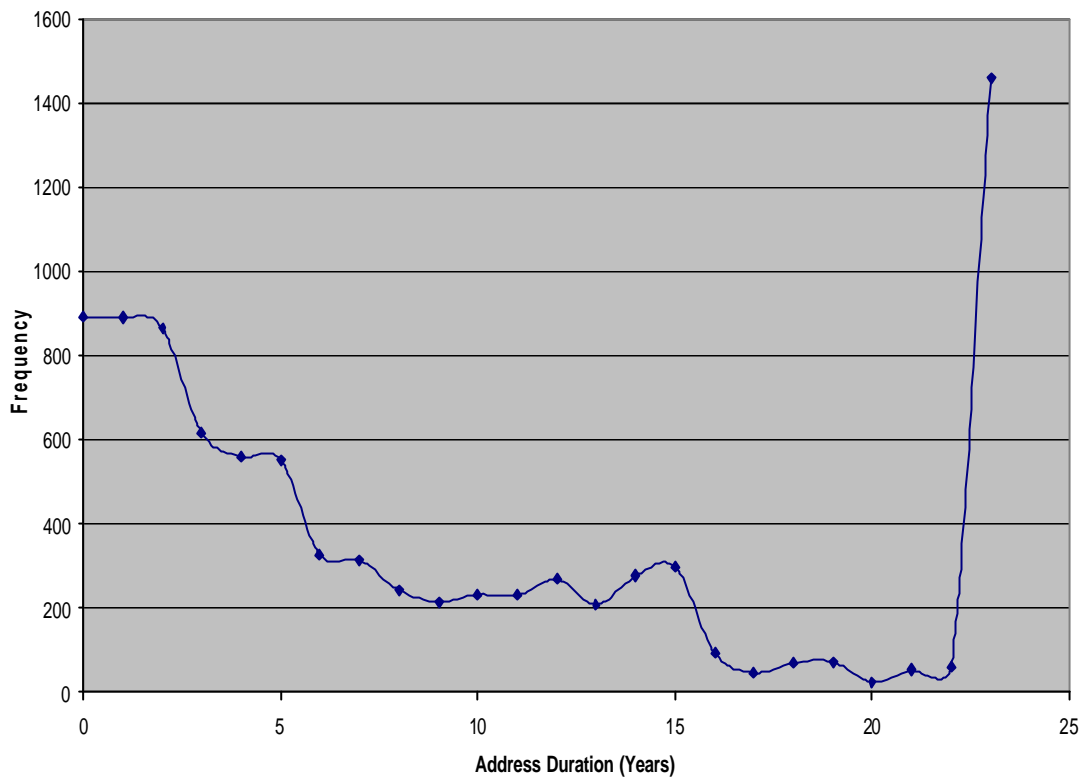
4.3.1 Change of address statistics can be derived in a number of different ways. The method chosen follows the above analysis on record currency and reveals statistics about current addresses. Some preliminary analysis has also been carried out on historical addresses.

4.3.2 Figure 22 shows the distribution of address duration for current records in the s8/s9 demographics. Because these demographics are based on date of birth, the people concerned are only in their early to mid thirties, and the maximum duration at an address is therefore only 15 years. Otherwise, the data shows the same peak for address duration of under 1 year as seen for DVLA. The bulge at years 5 (1999), 7 (1997) and 14 (1990) are taken to be effects of real events on peoples inclination to move. Similar, although not exactly correlated effects are seen in the DVLA data in Figure 6.

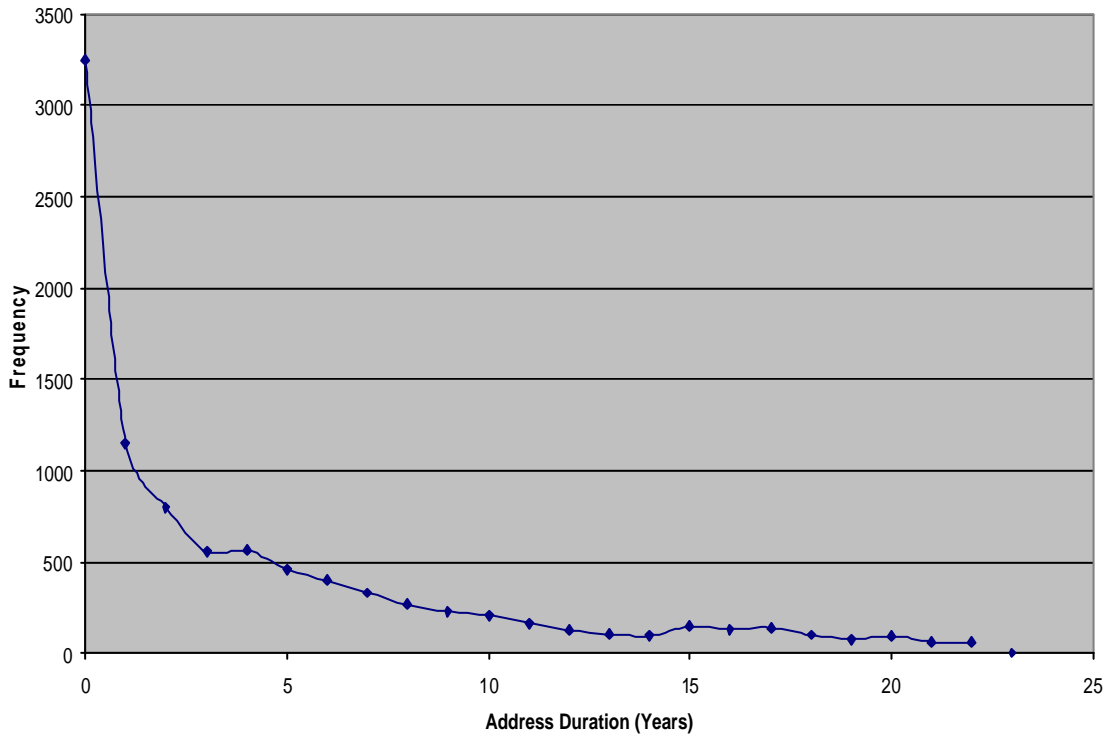


**Figure 22 - HMRC s8-s9 Current Address Duration**

- 4.3.3 For comparison, an analysis has been carried out on s1/s2, for which full address histories are also available. Only data for the current name has been plotted in Figure 23, to ensure the frequencies are not skewed by the presence of multiple name records. This data extends back to 1981, when the NIRS system was first introduced. We suppose that people whose current address Start Date pre-dates NIRS are held in the data with a start date of 1981, which explains the peak at year 23. These people have lived at their current address for more than 23 years. Surprisingly the peak for year zero is flattened: comparison with Figure 22 and Figure 6 (DVLA) would lead us to expect a peak of several thousand addresses. The data has been checked, and there is currently no explanation for the absence of this peak. It may be a peculiarity of the s1/s2 demographics during the last two years. Otherwise, peaks and troughs are seen in the rest of the data similar but not identical to that seen in Figure 22.
- 4.3.4 Finally an analysis of historical addresses, i.e. those for which Address End Date is set, is shown in Figure 24. Because these are historical address durations, a given record is not located in time in the same way as for the current duration data. For example, for a record that shows an address duration of 5 years, the Address Start Date could be anywhere from 1981 to 1998, whereas for the current duration data, an address duration of 5 years implies the Address Start Date was 5 years ago today. For this reason, the various peaks and troughs indicative of various market forces seen in the current duration data is smoothed out in the historical data. We do however see the characteristic peak for less than one year, strangely absent in Figure 23 (note that Figure 23 and Figure 24 covers the same demographic).



**Figure 23 - HMRC s1-s2 Current Address Duration**



**Figure 24 - HMRC s1-s2 Historical Address Duration**

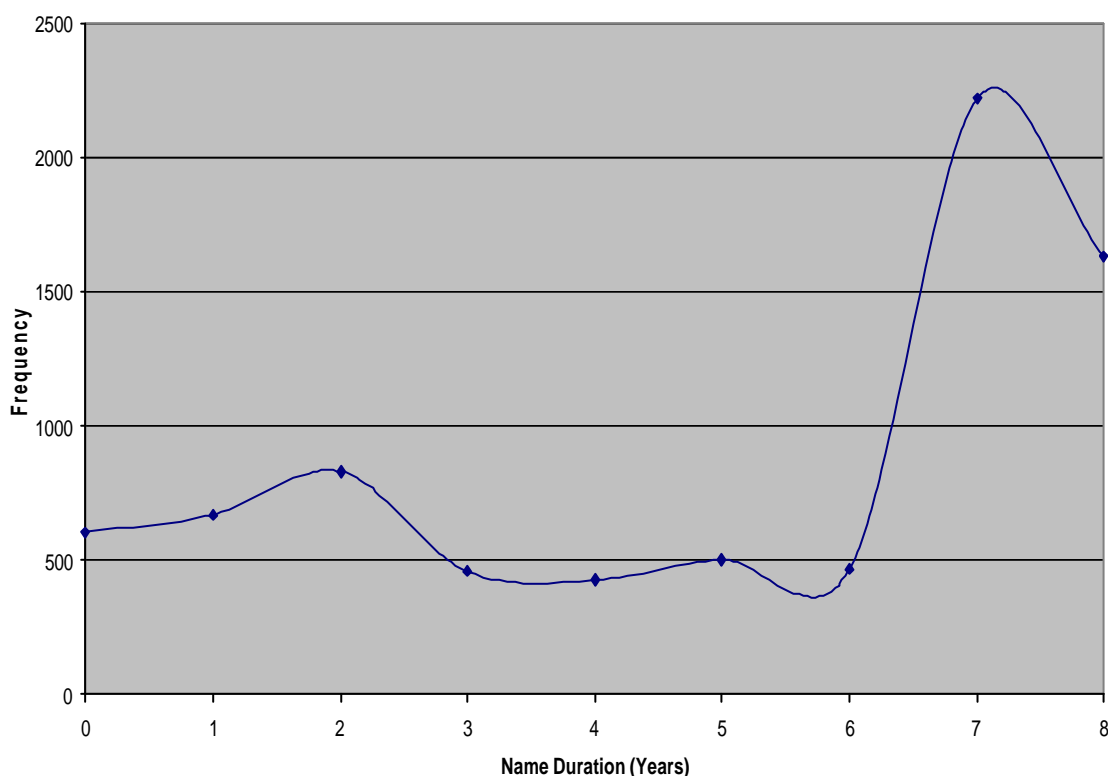
4.3.5 Average time at an address is derived from the underlying data as follows:

HMRC s8-s9 Current	3.7 years
HMRC s1-s2 Current	9.2 years
HMRC s1-s2 Historic	4.7 years

4.3.6 The reason s8-s9 is lower than s1-s2 is because of the younger age of these demographics. S1-s2 is assumed to be the more representative of the population as a whole. The average historic address duration is expected to be lower than the current address average, because it only covers people who have moved in the last 23 years. It does not cover people who have lived at an address for longer than 23 years, and is therefore lower than it otherwise might be. If the database went back further, say 80 years, then we would expect the current and historic averages to be roughly the same (subject to change in lifestyle trends over the period).

## 4.4 Change of name statistics

4.4.1 Change in name statistics have been derived in a similar way as for change in address. The s8-s9 demographics were used again for this purpose, as they contain the full name and address history. The name duration data is plotted in Figure 25.



**Figure 25 - HMRC s8-s9 Current Name Duration**

4.4.2 The data shows a more-or-less flat distribution of name duration up to 7 years ago. The peak at 7 and 8 years is an artefact of the database, in that the Name Start Date is only populated from December 1996 (8 years ago). This peak therefore represents the majority of people who have never changed their name.

- 4.4.3 A source of error in Name Duration statistics has been observed in the data in that a large number of Name records do not represent a genuine change in name. An additional name record appears to be created whenever new information about a persons name is recorded, such as a revised spelling.

## 5. Identity duplication

### 5.1 Date of birth, name and address matching

- 5.1.1 The occurrence of duplicate records for HMRC is 0.03% or 67 families with 67 member records out of the 207,882 split records.
- 5.1.2 Post QAS processing, the number of matches dropped to 66 family groups. This was because an address was cleansed which resulted in a match not being made. On inspection, this was ascertained to be an incorrect match in the pre-QAS data, and therefore 66 is taken as the more accurate match rate.
- 5.1.3 In summary, therefore, there are 66 duplicated HMRC records. These appear to be genuine cases of an individual holding two NINOs. The total number of HMRC combined records is 93,583, so the underlying duplication rate is 66 people out of (93,583 – 66) people, or 0.071%.
- 5.1.4 Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of

$$0.071\% \pm 2sd =$$

$$\mathbf{0.071\% \pm 0.017\%}$$

- 5.1.5 In a database of 40 million people, this equates to between 21,600 and 35,200 people with duplicate (i.e. two) records.

## Appendix E: Data assessment - UKPS

---

## 1. Analysis of dataset

1.1.1 The UKPS data sample holds c. 24,500 individual records.

### 1.2 Coverage

1.2.1 Overall coverage for UKPS PASS appears to be about 40 - 50% of the population. The data covers a period from 1998, i.e. it has been in operation for about 6 years, and given a 10-year renewal period for the passport, no more than 60% of the population would currently be expected within the database. The implication is that 44/60<sup>ths</sup> of the population are passport holders.

- Unusually low coverage is seen for s5 Scotland, suggesting lower numbers of passport holders in this region than the national average.
- The s9 demographic contains default dates.

### 1.3 Field analysis

1.3.1 The data comprised the following key fields:

- Unique reference number (Passport Number)
- Date of Birth
- Gender
- Place of Birth
- Name with separate Surname and Forename fields
- Address with separated Address fields
- Account Creation Date
- Last Update Date

1.3.2 In addition, the following fields were included:

- Former Name
- Second Address fields

1.3.3 Analysis of Date of Birth information shows good coverage of adults from age 16, and an increasing coverage of young children. There does appear to be a significant use of 1<sup>st</sup> January as a default birth date. In the full dataset, we would expect 33% +/- 24% of all 1<sup>st</sup> January dates to be defaults.

### 1.4 Identity duplication

1.4.1 Comparing name, address and date of birth, a number of people were found who appear to have two or more passport records. This is to be expected for renewals, legitimate duplicates, etc. 406 such records were found (1.68%).

1.4.2 Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of

More than one record: 1.68% +/- 2sd = **1.68% +/- 0.08%**

Two records: 1.65% +/- 2sd = **1.65% +/- 0.08%**

Three records: 0.03% +/- 2sd = **0.03% +/- 0.01%**

1.4.3 In a database of 10 million people, this equates to:-

More than one record between 160,000 and 176,000 people

Two records between 157,000 and 173,000 people

Three records between 20,000 and 40,000 people

## 1.5 Demographic analysis

1.5.1 Demographic differences derived by Individual Column Analysis were not particularly expected. However, demographic differences are apparent for Gender.

1.5.2 For Gender overall, the data contains 48% males, which extrapolates to the full dataset as 48.0% +/- 0.3%. However, significant variation to this was seen for s9 (DoB 1/1), where the rate drops to only 41% +/- 1.4%. Note that this is in contrast to the other datasets which show a predominantly male ratio in the s9 demographic.

## 2. Data structure

2.1.1 UKPS sent the data containing all nine demographics in a single file. The file was comma separated with data values double quoted. The format corresponded almost exactly with the specification and comprised the following fields:

- Unique reference number (Passport Number)
- Date of Birth
- Gender
- Place of Birth
- Name with separate Surname and Forename fields
- Address with separated Address fields
- Account Creation Date
- Last Update Date

2.1.2 In addition, the following fields were included:

- Former Name
- Second Address fields

### 3. Statistical summary

3.1.1 The UKPS datafile contains 24531 records. Three records with fields containing double quoted strings failed on the initial load but once the two pairs of double quotes around quotes were removed the records loaded successfully.

3.1.2 The number of records per demographic are as below:

3.1.3 s1 – Typical Dataset by name

	Frequency	Percentage s0
Records in sample	1766	7.2

3.1.4 s2 – Typical Dataset by name

	Frequency	Percentage s0
Records in sample	1456	5.9

3.1.5 s3 – Typical suburban dataset by geographic area (postcode and area name)

3.1.6 Sample based on the postcode or area name of a typical suburban area. Includes some records with incorrectly formatted postcodes for different postal areas. These occurrences are when AAN N\*\* postcodes have been read as AANN \*\*\*. The extract has included all records where the name of the area sampled appears in an address column. This results in records containing the area name in a road name e.g. "Oxford Road" or house name e.g. "Ullswater Cottage" which are not necessarily in the area that was intended to be sampled. The exact numbers of records which fall outside the intended criteria of the sample will be calculable after further analysis during data cleansing.

	Frequency	Percentage s0
Records in sample	5384	21.9
String in other address	uncalculated	

3.1.7 s4 – Covers name issues and address issues on houses that have been converted into flats. (postcode)

3.1.8 Postcode sample of an area which has a high number of addresses of flats in converted houses. With the sample criteria being just a postcode and not an area name, the data appeared more consistent than s3.

	Frequency	Percentage s0
Records in sample	5759	23.5
Postcode only in c/o Address	1	0
Invalid format postcode	1	0

3.1.9 s5 – Covers a rural area in Scotland (postcode)

3.1.10 This Scottish postcode sample is showed a high consistency in the data. Only one postcode was invalid and that still fell rightfully within the sample. One record in the sample didn't have the correct postcode but did have a c/o address with the postcode criteria.

**Citizen Information Project**

	Frequency	Percentage
Records in sample	1555	6.3
Postcode only in c/o Address	1	0
Invalid format postcode	1	0

3.1.11 s6 – Covers issues around Welsh names and addresses (postcode and area name)

3.1.12 Sample 6 was fairly accurate for a postcode sample and has a number of records within the estimate of 2000 -3000. There were 5 occurrences of the area string appearing in address fields in areas outside of the intended sample, three of these records did not contain the full string thus implying only part of the string was used in the extract. e.g. If the string was “Ports” seeking Portsmouth addresses the result set would include Port Talbot and Portadown addresses too.

	Frequency	Percentage s0
Records in sample	2037	8.3
Area string in outside address	5	0

3.1.13 s7 – Covers issues related to high density urban areas and high rise flat blocks

3.1.14 This ANN N\*\* format postcode sample has 163 records which have the postcode format AANN N\*\*. This postcode is for an area well outside of intended sample.

	Frequency	Percentage s0
Records in sample	4240	17.2
AANN N** format postcodes	163	4
Invalid format postcode	1	0
Despatch address	1	0

3.1.15 s8 – Typical dataset by date of birth

3.1.16 Demographic sample for people with a birth date such as 23/02/1973, with the format CCYYMMDD (19730223).

	Frequency	Percentage s0
Records in sample	1064	4.3

3.1.17 s9 – Covers issues around nominated date of birth being 1<sup>st</sup> January

3.1.18 Demographic sample for people with a birth date of the first of January for a chosen year such as 01/01/1976 with the format CCYYMMDD (19760101). There are 19% more records of this date than in demographic s8 due possibly to birthdates being recorded as the first of the year when the exact date is unknown.

	Frequency	Percentage
Records in sample	1270	5.2

## 4. Individual column analysis

### 4.1 Passport number

- 4.1.1 This id is unique and not null for each record. It should be noted that people can appear multiple times if they have been assigned more than one passport. This usually occurs when a previous passport is cancelled but can also happen when there is a need for person to have more than one passport for political reasons.
- 4.1.2 Two of the demographic samples intersect (i.e. contain some of the same records), as shown in the following table. For the avoidance of doubt, an intersection of 1 record indicates that 1 record appears twice, once in each demographic. An intersection of 2 indicates that 2 records appear twice, once in each demographic.

Intersected Demographic	Frequency
s2-s4	2
s2-s7	1

- 4.1.3 UKPS - ID Startdatetime, ID Enddatetime, Temp ref id, Temp Ref Startdatetime, Temp Ref Enddatetime
- 4.1.4 The above columns are null for all records.

### 4.2 Passport status

- 4.2.1 Status is either C (Cancelled) or I (Issued) and there were no nulls. People with a cancelled passport are likely to have been issued a new passport and thus will probably appear more than once in the sample.

Value	Format	Frequency	Percentage s0
<b>Cancelled</b>	<b>C</b>	<b>584</b>	<b>2</b>
<b>Issued</b>	<b>I</b>	<b>23947</b>	<b>98</b>

### 4.3 Date of birth

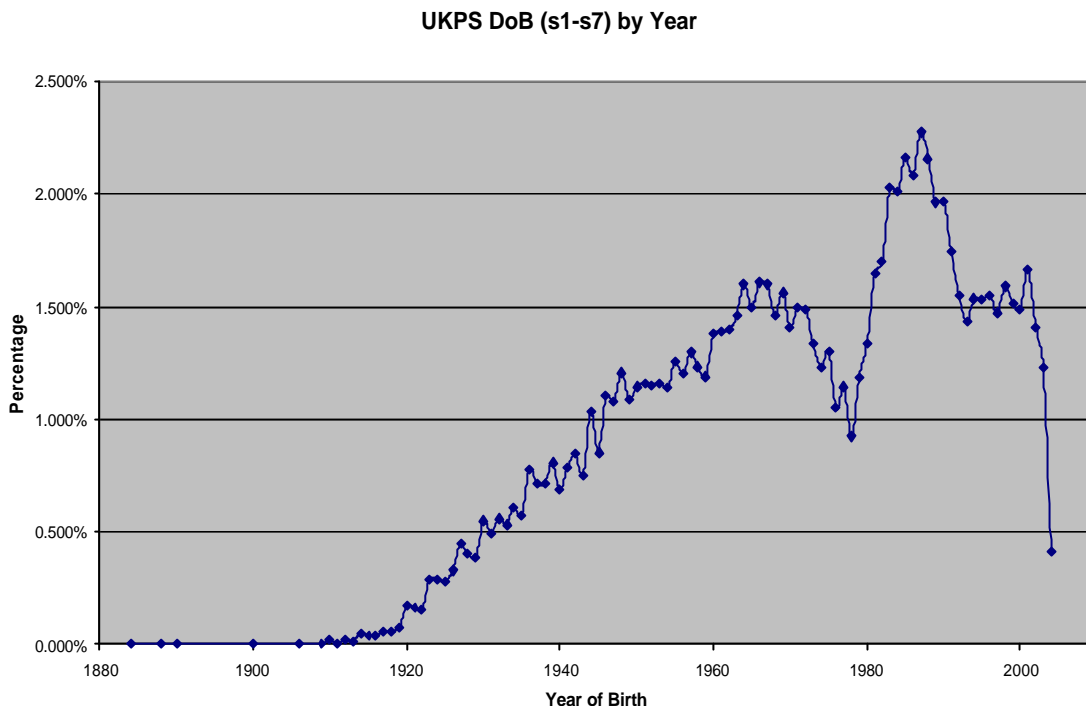
- 4.3.1 In the format CCYYMMDD there were no nulls. There were dates recorded for 1890, 1894 and 1900 which may be dubious. Apart from the expected high frequencies around the s8 and s9 sample birthdates no date appeared grater than 8 times.

Value	Format	Frequency	Percentage s0
<b>Null</b>	<b>CCYYMMDD</b>	<b>0</b>	<b>0</b>
<b>not null</b>		<b>24531</b>	<b>100</b>
< 1890		3	0

- 4.3.2 The UKPS DoB data, excluding s8 and s9, is plotted in Figure 26 and ranges from the late 1800's right up to 2004 (22197 values). In fact, a significant proportion of applications are for newborn babies (see Figure 27), and so unlike DVLA and IR, UKPS is a source of data for under 16's, as is of course GRO.

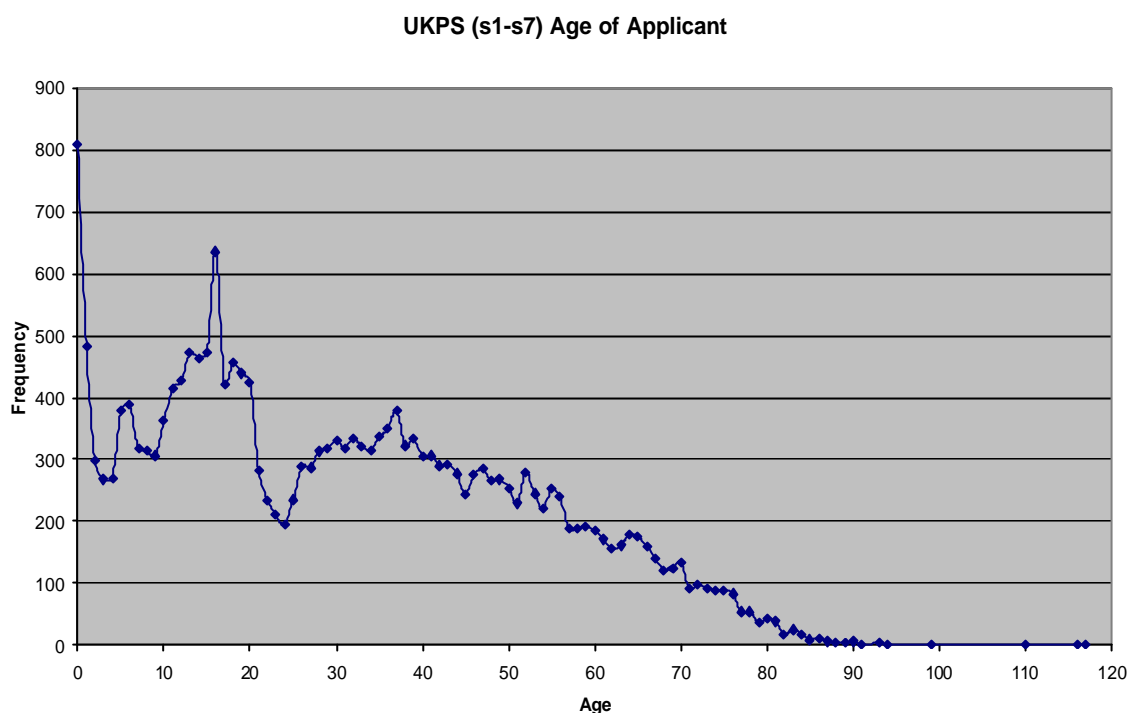
**Citizen Information Project**

There is an unexpected peak around 1988 possibly due to children being required to have their own passport for 1998 (most of the peak would be children 10-16 years old i.e. birth date of 1988). The dip in 1979 is primarily a feature of general birth rate.



**Figure 26 – UKPS DoB by Year (s1-s7 only)**

4.3.3 Figure 27 shows the age of each applicant, as calculated by subtracting DoB from Account Creation Date. As already mentioned, there is a significant peak for newborn, who may be issued with a five-year passport, and a corresponding peak for 5-6 year olds indicating passport renewal. There is also a peak for 16 year olds applying for passports, in support of the above interpretation for the DoB peak around 1988.



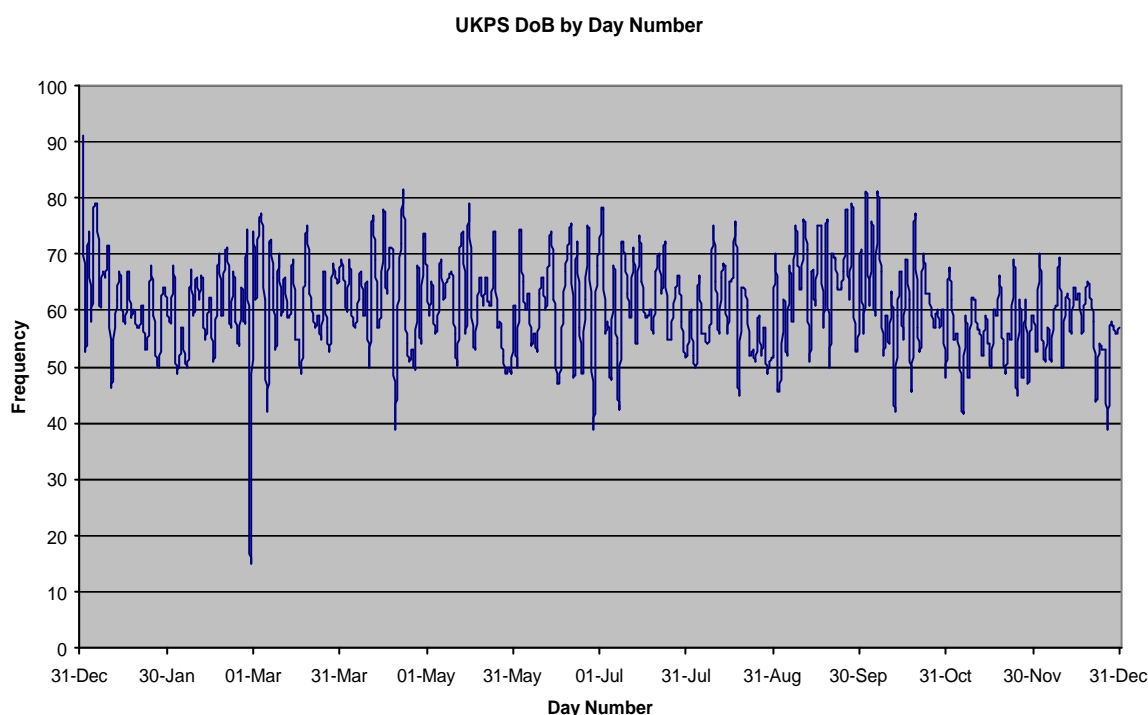
**Figure 27 – Age of UKPS Applicant (s1-s7 only)**

- 4.3.4 The s1-s7 DoB data has been plotted again in Figure 28, this time to show frequency of particular days in the year. This is to indicate if there is a particular peak at 1<sup>st</sup> Jan in any year. The year data was removed, the remaining month-day information sorted, and the occurrence of each of the 366 values counted and plotted in Figure 28. (To simplify the process, the month-day data was combined with the year value 2004 (a leap year) to assist the production of the x-axis.)
- 4.3.5 Note the peak at 1<sup>st</sup> January and dip at 29<sup>th</sup> February. The average frequency for any given date is 60.7, and the standard deviation (sd) (ignoring data for 1/1 and 29/2) is 8.4 (14%) – nearly all the data falls within 1 sd, and most of it within 2. There is nothing else outside 3 sd. The theoretical average and sd for a probability of 1/365 (0.0027) and sample size as specified are 60.8 and 7.8 respectively – very close to the measured values.
- 4.3.6 The peak at 01/01 is 91, 1.50 times the average value, and some 4 sd larger. This is very unlikely to be random and we assume it is a real feature of the data, suggesting that approximately two thirds of the 01/01 births are default values. The measured probability of a date being 01/01 is  $(91 / 22197) = 0.0041$ . The theoretical sd for this value is 9.5 (10.5%), and so in the full population, we might expect an occurrence rate of  $0.0041 \pm 0.0009$  (2sd) (i.e. within 21%). We can calculate the uncertainty in the 1.50 ratio (1<sup>st</sup> Jan occurrence over average occurrence) by combining the statistical uncertainties in each term (i.e. combining 14% and 10.5%). This is done using the standard “sum of squares” formula to produce an sd of about 18% (of the value of the ratio) or an error range of  $1.50 \pm 0.26$  (1sd). We do not know the full population size, but we do know the s8 and s9 sizes for the full population, i.e. 1064 and 1270 respectively, and the ratio

**Citizen Information Project**

s9/s8 should be similar to 1.50. In fact, taking into account the variation in birth rate between the two demographics, the ratio is derived as 0.7 / 0.5 or 1.42. The two values (1.50 and 1.42) are hence well within 1sd of each other, and therefore consistent.

- 4.3.7 The corollary is that  $1/1.50 = 67\% \pm 24\%$  (2sd) of 1<sup>st</sup> January dates are in fact correct, whereas  $33\% \pm 24\%$  are default dates.
- 4.3.8 The dip at 29/02 occurs because only 1 year in four is a leap year, and hence we would expect the number of occurrences of 29/02 to be approximately  $\frac{1}{4}$  of the average, i.e. 15. There are in fact 15 occurrences.



**Figure 28 – UKPS DoB by Day of Year**

## 4.4 Miscellaneous fields

### Verified date of birth, date of death, marital status

- 4.4.1 The above columns are null for all records.

## 4.5 Gender

- 4.5.1 Gender is recorded as M (Male) and F (Female) and there were no nulls.

Value	Format	Frequency	Percentage s0
null		0	0
Female	F	12760	52.0
Male	M	11771	48.0

## 4.6 Place of Birth

4.6.1 No null values. Place of birth is recorded as a town or less frequently a country.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>24531</b>	<b>100</b>

## 4.7 Name elements

### Title

4.7.1 Whilst nearly 99% of titles were of the five most regularly formats there were odd cases of military or religious titles. The invalid titles included numbers or in four records just a dot.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>138</b>	<b>0.6</b>
<b>not null</b>		<b>24393</b>	<b>99.4</b>
Mr		10750	43.8
Miss		6264	25.5
Mrs		5653	23.0
Ms		813	3.3
Master		787	3.2
Dr		58	0
Invalid		11	0

### Forename(s)

4.7.2 Includes multiple forenames.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>2</b>	<b>0</b>
<b>not null</b>		<b>24529</b>	<b>100</b>

### Surname

4.7.3 No records have a null surname.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>24531</b>	<b>100</b>

### Former Name

4.7.4 Former name is null in 18592 records, 10 further records have and invalid entry and are recorded as 13 NA (not applicable)

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>18595</b>	<b>75.8</b>
<b>not null</b>		<b>5936</b>	<b>24.2</b>
Not applicable	NA	13	0
Invalid (dot or dash)		10	0

## 4.8 Address elements

### Address 1 Address line 1

- 4.8.1 First line of address commonly containing flat or house name or number and street name.

Value	Format	Frequency	Percentage s0
null		2	0
not null		24529	100

### Address 1 Address line 2

- 4.8.2 Second line of address usually containing street name where not in address line 1 or area name.

Value	Format	Frequency	Percentage s0
null		9706	39.6
not null		14825	60.4

### Address 1 Address line 3 (Town)

- 4.8.3 Third line of address containing town name.

Value	Format	Frequency	Percentage s0
null		626	2.6
not null		23905	97.4

### Address 1 County

- 4.8.4 Country values UK, GB, Wales, England and others were found in this county column.

Value	Format	Frequency	Percentage s0
null		5129	20.9
not null		19402	79.1
Country		Approx 400	1.6

### Address 1 Postcode

- 4.8.5 There were invalid postcodes which did not meet accepted postcode formats.

Value	Format	Frequency	Percentage s0
null		249	1.1
not null		24282	98.9
Invalid		90	

### Address Verification

- 4.8.6 Null for all records.

### Address Match Level

- 4.8.7 PAF match level.

**Citizen Information Project**

Value	Format	Frequency	Percentage s0	Percentage non null
<b>Null (Unmatched)</b>		<b>18786</b>	<b>76.6</b>	
Full Match	F	1963	8.0	34%
Update Full Match	G	780	3.2	14%
Not Matched	N	1237	5.0	21%
Partial Match	P	1260	5.1	22%
Update Partial Match	Q	502	2.1	9%
Unmatched	U	3	0	0%

4.8.8 Note that UKPS only started using QAS Batch in c. May 2003, which accounts for the high null rate: i.e. all data pre May 2003 will be null.

4.8.9 UKPS sets the Full Match if there is a PAF match code of B, C, D or K (See Lot 2, ref [12]), and if there is a match on electoral roll surname. Otherwise, if surname does not match a Partial Match code is set. We understand that UKPS achieves c. 75% match rate across all their current data. Of the non-null values listed above, 56% match (Full Match plus Partial Match), and 79% match following update. The results obtained in Lot 2 indicate 95% match against PAF.

**Address 2 Address line 1**

4.8.10 Delivery address line 1. Two records were of a date value.

Value	Format	Frequency	Percentage s0
<b>Null</b>		<b>24377</b>	<b>99.4</b>
<b>not null</b>		<b>154</b>	<b>0.6</b>
Contained date		2	0

**Address 2 Address line 2**

4.8.11 Delivery address line 2. Two records were of a time value.

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>24379</b>	<b>99.4</b>
<b>not null</b>		<b>152</b>	<b>0.6</b>
Contained time		2	0

**Address 2 Address line 3 (Town)**

4.8.12 Delivery address line 3 usually containing town name.

Value	Format	Frequency	Percentage s0
<b>Null</b>		<b>24396</b>	<b>99.4</b>
<b>not null</b>		<b>135</b>	<b>0.6</b>

**Address 2 County**

4.8.13 Mostly containing county

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>24453</b>	<b>99.7</b>
<b>not null</b>		<b>78</b>	<b>0.3</b>
postcode		4	0

### Address 2 Postcode

#### 4.8.14 Delivery address postcode

Value	Format	Frequency	Percentage s0
null		24390	99.4
not null		141	0.6

### 4.9 Last Update Date

Value	Format	Frequency	Percentage s0
Null		0	0
not null	CCYYMMDD	24531	100
1998		37	
1999		1393	
2000		2223	
2001		4633	
2002		5440	
2003		5225	
2004		5580	
Total		24531	

### UKPS Update Date by Year

Number of Records

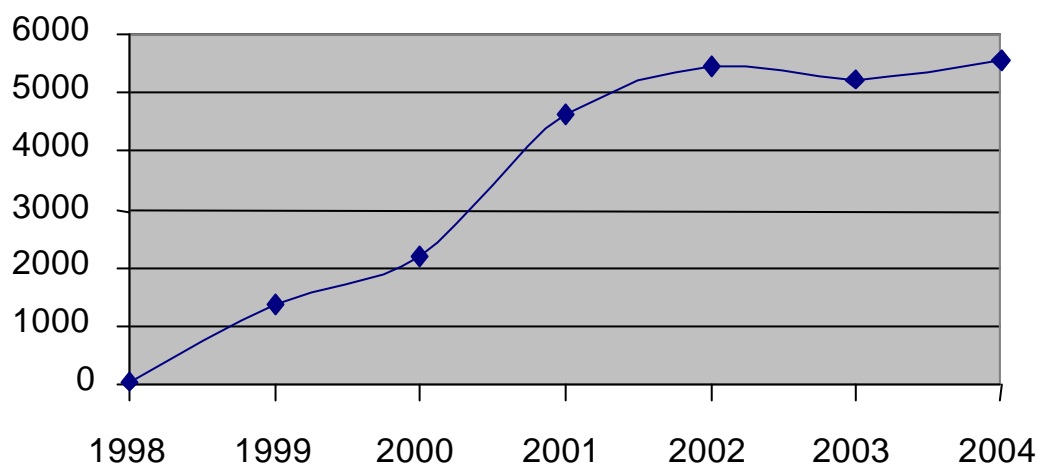


Figure 29 – UKPS Last Update Date by Year

Note that the update date is very close to the Creation Date, and the shape of the graph is similar to Figure 30.

### Last Update Time

Value	Format	Frequency	Percentage s0
null		0	0
not null	nnn or nnnn	24531	100

**Last Update User**

Value	Format	Frequency	Percentage s0
<b>Null</b>		<b>0</b>	<b>0</b>
<b>not null</b>		<b>24531</b>	<b>100</b>

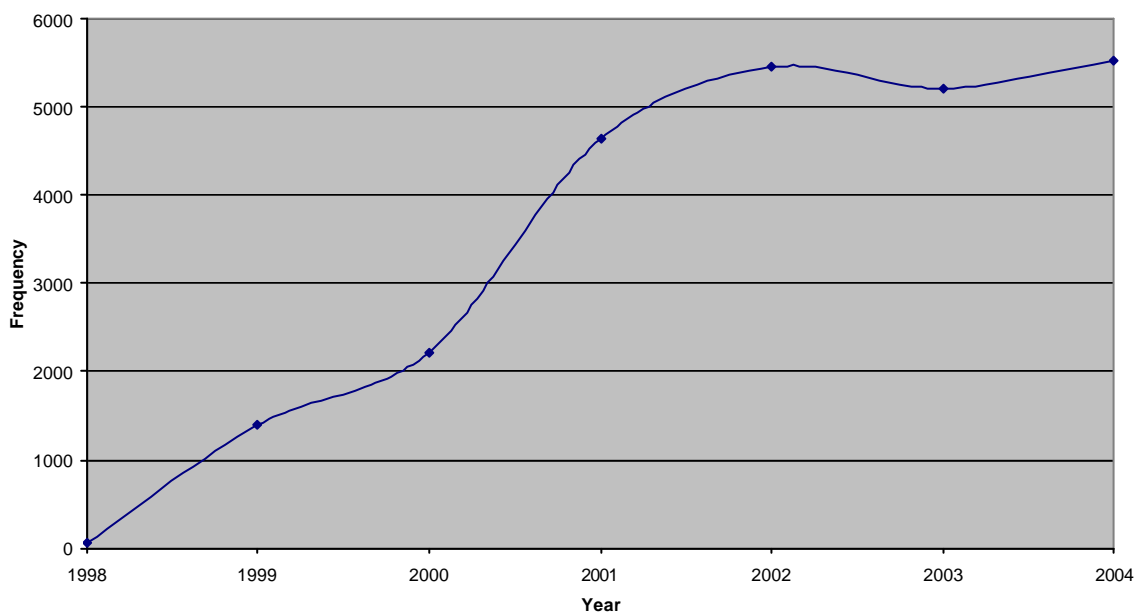
**Last Update User Location**

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>87</b>	<b>0.4</b>
<b>not null</b>		<b>24444</b>	<b>99.6</b>

**4.10 Account creation date**

Value	Format	Frequency	Percentage s0
<b>Null</b>		<b>0</b>	<b>0</b>
<b>not null</b>	<b>CCYYMMDD</b>	<b>24531</b>	<b>100</b>
1998		50	
1999		1399	
2000		2225	
2001		4644	
2002		5457	
2003		5218	
2004		5538	
<b>Total</b>		<b>24531</b>	

**UKPS Account Creation Date by Year**



**Figure 30 – UKPS Account Creation Date by Year**

**Last Update User Location**

Value	Format	Frequency	Percentage s0
<b>null</b>		<b>0</b>	<b>0</b>
<b>not null</b>	<b>CCYYMMDD</b>	<b>24531</b>	<b>100</b>

## Office of Creation

4.10.1 Office of creation may relate to location of persons within sample.

Value	Format	Frequency	Percentage s0
Belfast		354	1.4
Durham		5780	23.6
Glasgow		1591	6.5
Liverpool		1946	7.9
London		1121	4.6
Newport		8357	34.1
Peterborough		5382	21.9

## 4.11 PASS document type

4.11.1 Multiple values of verification documents e.g. LBC, SBC, PPT etc.

Value	Format	Frequency	Percentage s0
Null		2	0
not null		24529	100

## 5. Identity duplication

### 5.1 Date of birth, name and address matching

5.1.1 The headline match rate for UKPS is 1.35%, or 406 family groups with 414 records out of 30,647 split records. This is higher than the other datasets, but UKPS business rules permit multiple records: each new passport application will exist in the database as a separate record.

5.1.2 There was no difference in match rate after QAS address cleansing, however, inspection of the data reveals that post QAS address cleansing, in fact one false match had dropped out, and a new match had come in.

5.1.3 There are some groups with more than two records. The distribution is in fact as follows:

2 records 398

3 records 8

5.1.4 In summary, the total number of individual people records is  $(24,531 - 406) = 24,125$ , and the number of people with more than one record is 406 or 1.68%. There are 398 people (1.65%) with two records and 8 people (0.03%) with three records.

5.1.5 Extrapolating to the full database, we might expect an occurrence rate, within a 95% confidence band of:

More than one record:  $1.68\% \pm 2sd = 1.68\% \pm 0.08\%$

Two records:  $1.65\% \pm 2sd = 1.65\% \pm 0.08\%$

Three records:  $0.03\% \pm 2sd = 0.03\% \pm 0.01\%$

5.1.6 In a database of 10 million people, this equates to:-

More than one record	between 160,000 and 176,000 people
Two records	between 157,000 and 173,000 people
Three records	between 20,000 and 40,000 people