The Interplay among Censuses, Surveys and Administrative Data

John `Mac’ McDonald

Sunday, 7 March 2010
Social Statistics: The Interplay among Censuses, Surveys and Administrative Data

Symposium 99 - Combining Data from Different Sources
The Symposium will be titled “Social Statistics: The Interplay among Censuses, Surveys and Administrative Data”. Members of the community, such as those from private organizations, governments, or universities, are invited to attend, particularly if they have a special interest in statistical or methodological issues resulting from the use of multiple sources of data (censuses, sample surveys or administrative data).

The first day will consist of workshops, while the following days will consist of both plenary and parallel sessions covering a variety of topics. Additional research and results will be presented via poster sessions.

We are soliciting contributed papers related to the methodological aspects of using multiple sources of data. Topics may include:

- Sampling Frames and Sample Design
- Coordinating Samples
- Content and Questionnaire Design
- Data Collection Methods and Acquisition of Administrative Data
- Supplementing Survey Data with Administrative Data
- Administrative Data for Direct Estimation
- Statistical Databases from Administrative Data (e.g., Population Registers)
- Imputation
- Weighting and Estimation
- Dissemination and Data Access
- Record Linkage Techniques
- Record Linkage Software
- Measurement Errors
- Response Burden
- Treatment of Nonresponse
- Confidentiality, Privacy and Ethical Issues
- Small Area Estimation

the methodological aspects of using multiple sources of data
- Sampling Frames and Sample Design
- Coordinating Samples
- Content and Questionnaire Design
- Data Collection Methods and Acquisition of Administrative Data
- Supplementing Survey Data with Administrative Data
- Administrative Data for Direct Estimation
- Statistical Databases from Administrative Data (e.g., Population Registers)
- Imputation
- Weighting and Estimation
- Dissemination and Data Access
- Record Linkage Techniques
- Record Linkage Software
- Measurement Errors
- Response Burden
- Treatment of Nonresponse
- Confidentiality, Privacy and Ethical Issues
- Small Area Estimation
15th GSS Methodology Conference, 1 July 2010

Making data work harder
Methodology sessions on

- Introducing efficiencies
- Methodological change
- Innovation in methods
- Maintaining quality
- **Use of administrative data**
- **Data sharing**
- Survey/Analysis methods
- General statistical methods
JW ‘Mac’ McDonald – CV

EDUCATION

1980  Ph.D. in Biostatistics
      University of Washington, Seattle

1972  M.A. in Demography
      University of California at Berkeley

1970  B.A. in Mathematics & Economics
      University of California at Los Angeles

EMPLOYMENT HISTORY

2007–present  Professor of Longitudinal Social Statistics
               Institute of Education, University of London

1984–2007  Lecturer, Senior Lecturer, Reader and
           Professor of Social Statistics
           University of Southampton

1980–1984  Research Officer, World Fertility Survey
           International Statistical Institute
census data

vital registration data

survey data

church records
Perspectives on Longitudinal Surveys

Randall J. Olsen

“Perhaps the greatest unexploited opportunity for survey projects lies in administrative data.”
• exploring how researchers should best use administrative datasets that have become available in the UK

• determining how researchers can enhance longitudinal survey data by exploiting available administrative data

• training and capacity building - short courses
• Introduction to Data Linkage
• The Value of Data Linkage for Research
• Data Linkage - Methodological and Statistical Issues
• Longitudinal Data Linkage
• Event History Analysis
• Longitudinal Data Analysis
• Introduction to the National Pupil Database

• Quantitative Analysis Using the National Pupil Database

• Using Administrative Data to Analyse the Impact of Policy Initiatives

• Enhancing Longitudinal Surveys by Linking to Administrative Data

• one new course annually
Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.


**EACH** person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

and is but a few pages in length for others. In the case of a stillbirth, the entire volume is but a single page.

The person retains the same identity throughout the Book. Except for advancing age, he is the same person. Thinking backward he can remember the important pages of his Book even though he may have forgotten some of the words. To other persons, however, his identity must be proven. "Is the John Doe who enlists today in fact the same John Doe who was born eighteen years ago?"

Events of importance worth recording in the Book of Life are frequently put on record in different places since the person moves about the world throughout his lifetime. This makes it difficult to assemble this Book into a single compact volume. Yet, sometimes it is necessary to examine all of an individual's important records simultaneously. No one would read a novel, the pages of which were not assembled. Just so, it is necessary at times to link the various important records of a person's life.

The two most important pages in the book of Life are the first one and the last one. Consequently, in the process of record linkage the uniting of the fact-of-death with the fact-of-birth has been given a special name, "death clearance."

**IMPORTANT OF ASSEMBLING THE BOOK OF LIFE**

There are many uses for the important records of each person, brought together as a whole. At times, even now, such a collection is of sufficient value that it is made at considerable cost in time and money. Usually, it is the individual who is made to do the work since he alone knows where his records are on file. It is much more difficult for any other person or organization to assemble the records of his life since no personal cross-index exists to lead one to all of a person's records. It is important to many people and organizations to be able to assemble this type of information easily and efficiently.

**IMPORTANCE TO THE INDIVIDUAL**

Sooner or later most of us need to prove facts about ourselves. The most frequent facts are concerned with birth. Sometimes other facts are needed. Many of these are recorded in the vital records of marriage, divorce, adoption, legitimation, change of name, death and presumption of death. However, it is


[1412]
Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

On record in different places since the person moves about the world throughout his lifetime. This makes it difficult to assemble this Book into a single compact volume. Yet, sometimes it is necessary to examine all of an individual's important records simultaneously. No one would read a novel, the pages of which were not assembled. Just so, it is necessary at times to link

Events of importance worth recording in the Book of Life are frequently put on record in different places since the person moves about the world throughout his lifetime. This makes it difficult to assemble this Book into a single compact volume. Yet, sometimes it is necessary to examine all of an individual’s important records simultaneously. No one would read a novel, the pages of which were not assembled. Just so, it is necessary at times to link
Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.

Chief, National Office of Vital Statistics, U. S. Public Health Service,

EACH person in the world creates a Book of Life. This Book starts
with birth and ends with death. Its pages are made up of the records of
the principal events in life. Record linkage is the name given to the process
of assembling the pages of this Book into a volume.

The Book has many pages for some
and is but a few pages in length for
others. In the case of a stillbirth, the
entire volume is but a single page.

The person retains the same identity
throughout the Book. Except for ad-
vancing age, he is the same person.
Thinking backward he can remember
the important pages of his Book even
though he may have forgotten some of
the words. To other persons, however,
his identity must be proven. “Is the
John Doe who enlists today in fact the
same John Doe who was born eighteen
years ago?”

Events of importance worth recording
in the Book of Life are frequently put
on record in different places since the
person moves about the world through-
out his lifetime. This makes it difficult
to assemble this Book into a single
compact volume. Yet, sometimes it is
necessary to examine all of an individual’s important records simultaneously. No one would read a novel, the
pages of which were not assembled. Just so, it is necessary at times to link

the various important records of a
person’s life.

The two most important pages in the
Book of Life are the first one and the
last one. Consequently, in the process
of record linkage the uniting of the fact-of-death with the fact-of-birth has
been given a special name, “death
clearance.”

IMPORTANCE OF ASSEMBLING THE
BOOK OF LIFE:

There are many uses for the impor-
tant records of each person, brought
together as a whole. At times, even now,
such a collection is of sufficient value
that it is made at considerable cost in
time and money. Usually, it is the
individual who is made to do the work
since he alone knows where his records
are on file. It is much more difficult
for any other person or organization to
assemble the records of his life since
no personal cross-index exists to lead
one to all of a person’s records. It is
important to many people and or-
ganizations to be able to assemble
this type of information easily and
efficiently.

IMPORTANCE TO THE INDIVIDUAL

Sooner or later most of us need to
prove facts about ourselves. The most
frequent facts are concerned with birth.
Sometimes other facts are needed.
Many of these are recorded in the vital
records of marriage, divorce, adoption,
legitimation, change of name, death and
presumption of death. However, it is

* Modified form of paper given before the joint
conference of the Vital Statistics Council for Canada
and the Dominion Council of Health, held in Ottawa,
Ontario, Canada, on May 10, 1946.
Record Linkage

HALBERT L. DUNN, M.D., F.A.P.H.A.

the various important records of a person’s life.

The Book has many pages for some and is but a few pages in length for others. In the case of a stillbirth, the entire volume is but a single page.

The person retains the same identity throughout the Book. Except for advancing age, he is the same person. Thinking backward he can remember the important pages of his Book even though he may have forgotten some of the words. To other persons, however, his identity must be proven. “Is the John Doe who enlists today in fact the same John Doe who was born eighteen years ago?”

Events of importance worth recording in the Book of Life are frequently put on record in different places since the person moves about the world throughout his lifetime. This makes it difficult to assemble this Book into a single compact volume. Yet, sometimes it is necessary to examine all of an individual’s important records simultaneously. No one would read a novel, the pages of which were not assembled. Just so, it is necessary at times to link clearance.

IMPORTANCE OF ASSEMBLING THE BOOK OF LIFE

There are many uses for the important records of each person, brought together as a whole. At times, even now, such a collection is of sufficient value that it is made at considerable cost in time and money. Usually, it is the individual who is made to do the work since he alone knows where his records are on file. It is much more difficult for any other person or organization to assemble the records of his life since no personal cross-index exists to lead one to all of a person’s records. It is important to many people and organizations to be able to assemble this type of information easily and efficiently.

IMPORTANCE TO THE INDIVIDUAL

Sooner or later most of us need to prove facts about ourselves. The most frequent facts are concerned with birth. Sometimes other facts are needed. Many of these are recorded in the vital records of marriage, divorce, adoption, legitimation, change of name, death and presumption of death. However, it is

We will exploit new record-linkage and data-mining technology to create linked representative samples of individuals and family groups from the censuses of 1860, 1870, 1900 and 1910 to the 1880 census.

Linked samples will provide new opportunities for researchers to carry out individual-level analyses of social and geographic mobility and family transitions.
UK Longitudinal Studies

• ONS (England & Wales) Longitudinal Study
  4 birthdays, 1% sample, c. 500,000

• Scottish Longitudinal Study
  16 + 4 birthdays, 5% sample, c. 274,000

• Northern Ireland Longitudinal Study
  104 birthdays, 28% sample, c. 500,000

• NB absolute numbers & relative numbers
A goal of record linkage is to join together two files 1. Record Linkage/File matching

The model contains some unknown probabilities. Let there be two files called file A and file B. A file consists of a series of comparisons reported in a vector. The Fellegi-Sunter (1969) procedure is as follows:

- For each pair of records 𝑎, 𝑏 agree on comparison 𝑘 \( \in \{1, \ldots, K\} \) the outcomes of a1 \( \gamma \in \{0, 1\} \), where 1 means true match and 0 otherwise. If the variables in file A and B have no agreement on a given comparison 𝑘, then they are declared nonmatching with weight of nonmatch being 0. Otherwise send for clerical review at pre-set error rates. The log of computation time and accuracy are concerns.

For each pair of records 𝑎, 𝑏 \( \in \{\text{file A, file B}\} \), the outcomes of a1 \( \gamma \in \{0, 1\} \), then

\[ a, b \]

\[ \gamma \]

\[ \{0, 1\} \]

\[ \in \]

\[ 0 \]

\[ \text{otherwise} \]

\[ \text{if records} \]

\[ \text{if} \]

\[ \text{otherwise} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{impossible} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]

\[ \text{if} \]
Confounder Z

X → Z → Y

Sunday, 7 March 2010
What variables to match on?

What variables to match on?

- IDs are subject to problems of survey item non-response and measurement error
- 5 linkages: respondent-supplied NINo and 4 linkages using different combinations of sex, name, address and DOB
- as many linkages were made using non-NINo matches as were made using NINo matches
- former also relatively accurate in terms of false-positive and false-negative linkage rates
Analysis of linked data

- Scheuren F & Winkler WE (1993) Regression analysis of data files that are computer matched - part 1. Survey Methodology 19, 39-58
Analysis of Linked Data

• What should the linker do to help the analyst?
• What should the analyst know about the linkage and how should that information be used?
• In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analysis done.
Quality of linkage

- 6.0% of deaths of residents of Northern Ireland could not be linked to a census record
- for those < 65 years at census, this linkage would exclude from analysis 20.2% of suicides and 19.7% of deaths by external causes
<table>
<thead>
<tr>
<th>Category</th>
<th>E&amp;W LS</th>
<th>SLS</th>
<th>NILS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Births to Sample Fathers</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Stillbirths/ Infant Deaths</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Internal Migration</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Cancer Registrations</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Hospital Episodes</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Marriages</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Claimant Count</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Migration flows

• increasingly important
• difficult to measure
• proxy admin data sources often used
• small area estimates of local pop change
• differentials by age, sex, LTLI, ...

• new and improved methods needed, e.g. Bayesian methods for estimates of precision, missing data, inconsistent data, calibration
References


### British residents born in New Commonwealth

<table>
<thead>
<tr>
<th>Residence in 1966</th>
<th>Residence in 1971</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>CC</td>
<td>118</td>
<td>12</td>
<td>7</td>
<td>23</td>
<td>160</td>
</tr>
<tr>
<td>L&amp;Y</td>
<td>L&amp;Y</td>
<td>14</td>
<td>2</td>
<td>127</td>
<td>86</td>
<td>130</td>
</tr>
<tr>
<td>WM</td>
<td>WM</td>
<td>8</td>
<td>69</td>
<td>2548</td>
<td>107</td>
<td>2732</td>
</tr>
<tr>
<td>GL</td>
<td>GL</td>
<td>12</td>
<td>110</td>
<td>88</td>
<td>7712</td>
<td>7922</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Total</strong></td>
<td><strong>152</strong></td>
<td><strong>2318</strong></td>
<td><strong>2729</strong></td>
<td><strong>7972</strong></td>
<td><strong>13171</strong></td>
</tr>
</tbody>
</table>

CC: Central Clydeside  
L&Y: Lancashire & Yorkshire  
WM: West Midlands  
GL: Greater London
Migration from 1966 to 1971 (flows)

<table>
<thead>
<tr>
<th>Residence in 1966</th>
<th>Residence in 1971</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>L&amp;Y</td>
</tr>
<tr>
<td>CC</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>L&amp;Y</td>
<td>14</td>
<td>86</td>
</tr>
<tr>
<td>WM</td>
<td>8</td>
<td>69</td>
</tr>
<tr>
<td>GL</td>
<td>12</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>191</td>
</tr>
</tbody>
</table>

Only marginal totals of flows known

<table>
<thead>
<tr>
<th>Residence in 1966</th>
<th>CC</th>
<th>L&amp;Y</th>
<th>WM</th>
<th>GL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>L&amp;Y</td>
<td>230</td>
<td></td>
<td></td>
<td></td>
<td>230</td>
</tr>
<tr>
<td>WM</td>
<td>184</td>
<td></td>
<td></td>
<td></td>
<td>184</td>
</tr>
<tr>
<td>GL</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td>210</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>191</td>
<td>181</td>
<td>260</td>
<td>666</td>
</tr>
</tbody>
</table>

CC: Central Clydeside  
L&Y: Lancashire & Yorkshire  
WM: West Midlands  
GL: Greater London
Take home points

• use NILS to study the ‘book of life’ (lifecourse)
• linkages with other data sources extremely valuable, e.g. to study/estimate migration flows
• 28% sample permits small area estimation at a low level of granularity
• conceptualize the linkage and analysis steps as part of a single statistical system
• linkage methodology should be transparent and quality of linkage studied