Statistics
New Zealand
TATAURANGA AOTEAROA

Documentation of the Process and Results of Linking Cancer Data with Census Data 1981 to 2001

Citation

Statistics New Zealand. (2008). *Documentation of the process and results of linking cancer data with census data 1981 to 2001.* Wellington: Statistics New Zealand

Published in July 2008 by

Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

ISBN 978-0-478-31536-3 (online)

Acknowledgements

This paper was prepared by Barb Lash, statistical analyst in the Social Conditions Business Unit, Statistics New Zealand. Nalin Patel, statistical analyst at Statistics New Zealand, assisted with the first draft of this paper by writing up the results of the linking for the 1981 to 1996 cohorts. June Atkinson, data manager at the Wellington School of Medicine and Health Sciences, Penny Barber, senior methodologist at Statistics New Zealand, and John Pearson from Statistics New Zealand, reviewed the paper and made helpful comments.

Standards and information

Liability statement

While all care and diligence has been used in processing, analysing and extracting data and information in this report, Statistics New Zealand gives no warranty it is error free and will not be liable for any loss or damage suffered as a result of the use, directly or indirectly, of information in this report.

Statistics New Zealand's Information Centre

For further help finding and using statistical information available on our website or in the INFOS database, contact the Information Centre:

Email: info@stats.govt.nz
Phone toll-free: 0508 525 525
Phone: (+64) (04) 931 4600
Fax: (+64) (04) 931 4610
Post: PO Box 2922, Wellington
Website: www.stats.govt.nz

Contents

Tables

Figures

1 Introduction

1.1 Purpose

The purpose of this project was to link all (or as many as possible) cancer records with the census record for the same person. The information available on the census records combined with the information available on the cancer records will provide a very rich source of information for research.

1.2 Overview

Figure 1 gives an overview of the process. This paper only documents the linkage part of the process, that is, the method used to link the files and the results of the linkage.

Since the census and cancer files do not contain names, probabilistic links were made on the basis of where the person lives and various demographic characteristics. The files used for linking only contain the variables used in the linking process, plus a unique identity number for each record.

Once the links have been made, the unique identity numbers are used to join the census files, with extra variables, and the cancer files, which also have extra variables. (This part of the project is not documented here.)

The project used both QualityStage and SAS software, whichever was appropriate software for a particular task. QualityStage is the software used in Statistics New Zealand to link data. SAS is good for producing reports, exploring and manipulating data. Figure 2 shows the stages of the process where SAS and QualityStage were used.

1.3 Cohorts

A cohort is a group of people experiencing the same event in the same period of time. In this project, a cohort for a census is the people with new cancer records in the period from the day after the census up to the day of the next census1 (inclusive). Five censuses (1981, 1986, 1991, 1996, and 2001) were used as a basis of five cohorts.

The cancer records to be linked with each census had the following characteristics:

- The person was alive on day of census.
- A new cancer record was created in the period from the day after the census up to the day of the next census (inclusive).
- If there was more than one cancer record for a person, only one record for that person was included in the data for linking. That one record included all information that could be used for linking from all records for that person.

The periods covered by the new cancer records are shown in the following table.

[1] Except for the 2001 Census cohort. Because cancer files were not available for the whole period, the period for the 2001 Census cohort ended on 31 December 2004.

Table 1.1 Periods covered by new cancer records 1981–2001

Census year for cohort	Date of census	Start of period for new cancer records	End of period for new cancer records
1981	24 March 1981	25 March 1981	4 March 1986
1986	4 March 1986	5 March 1986	5 March 1991
1991	5 March 1991	6 March 1991	5 March 1996
1996	5 March 1996	6 March 1996	6 March 2001
2001	6 March 2001	7 March 2001	31 December 2004

Figure 1.1 Overview of process

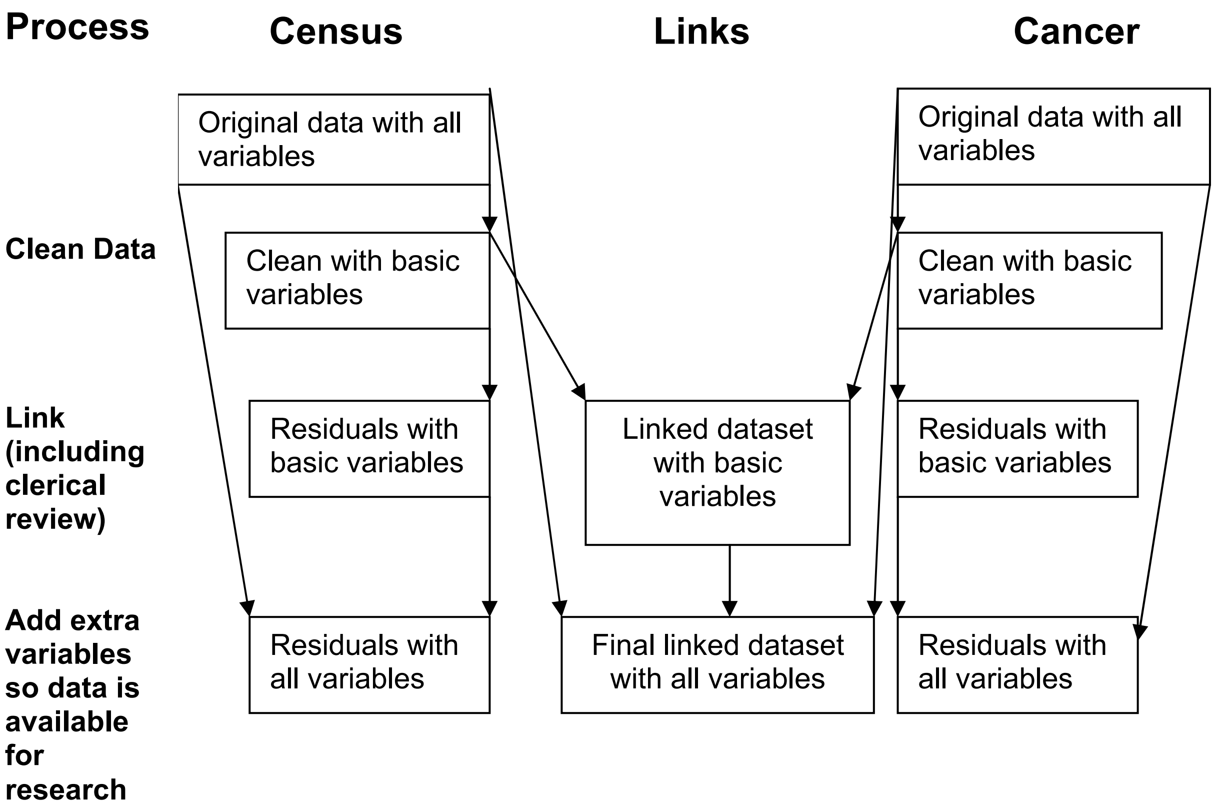
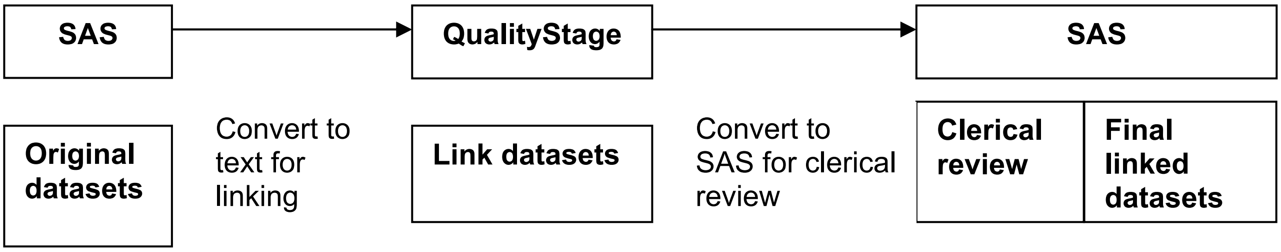

Process
Census
Links
Cancer
Original data with all variables
Original data with all variables
Clean Data
Clean with basic variables
Clean with basic variables
Link (including clerical review)
Residuals with basic variables
Linked dataset with basic variables
Residuals with basic variables
Add extra variables so data is available for research
Residuals with all variables
Final linked dataset with all variables
Residuals with all variables

Figure 1.2 Use of QualityStage and SAS



2 Description of files used for linking

2.1 Introduction

The census and the cancer files contained different variables which are listed below. In general, the formats of the variables on one file are the same as the corresponding variables on the other file. Extra variables on each file provide information for clerical review and the identity numbers for each record were used to add additional variables after the linking was complete (see Figure 1).

2.2 Country of birth and ethnicity adjustment factors

The cancer trends pilot study (Smith 2005 section 4.5) identified that the variables country of birth and ethnicity are not independent although the record linking model assumes that matching variables are independent. The two interactions between these variables that have the most impact on the linking are: Pacific People born in the Pacific Islands, and Asians born in Asia.

Two adjustment factors were added to each of the files (census and cancer) to lower the weight for records where the correlation between the two variables would increase the likelihood of the records linking when they were not a match. The adjustment factors consisted of constructing variables asianfix and pacfix on the census file, and asianfixw and pacfixw on the cancer file.

Asianfix and asianfixw had values for "yes" if the person was born in Asia and of Asian ethnicity, "no" if the person was either born in Asia or of Asian ethnicity, but not both, and "not applicable" if the person was neither born in Asia nor of Asian ethnicity. The variables Pacfix and pacfixw had values for "yes" if the person was born in the Pacific Islands and of Pacific ethnicity, "no" if the person either born in the Pacific Islands or of Pacific ethnicity, but not both, and "not applicable" if the person was neither born in the Pacific Islands nor of Pacific ethnicity.

2.3 Contents of census file

Not all of the records on the census dataset were eligible for linking as the population required for linking is the New Zealand usually resident population. The Technical Report and Feasibility Assessment of the Cancer Trends Pilot Study (Smith 2005) provides details of how the 3,516,513 census records from 1996 were selected to be included in the records available for linking.

> *The 1996 Census dataset contained 3,786,993 observations, although not all of these records were eligible for linkage. The population required for File A is the New Zealand usually resident population, which at the time was estimated to be 3,618,302; people overseas and absentees are excluded.*
>
> *Not all of these 3,618,302 records were available for linking. 101,789 individual records were added to the census dataset as 'dummies' to adjust for the census undercount. These records should obviously not be included in the record linkage and once they were removed, 3,516,513 records remained on File A. Table 2 shows the number of observations on the 1996 Census dataset and some key breakdowns.*

Table 2.1 Type of records on the 1996 census file

Group	*Number of People*
New Zealand Adult	*2,786,221 **
New Zealand Child	*832,081 **
Overseas Resident Working in NZ	*994*
Overseas Adult	*58,501*
Overseas Child	*3,749*
Absentee	*105,447*
Total	*3,786,993*

** Including 101,789 dummy records*

*** File A contains: NZ Adults + NZ Children - Dummy Records (or 3,516,513 records)*

(Smith 2005 page 5)

Records for the other census years were selected in a similar manner.

Table 2.2 Number of records on the census linking files 1981–2001

Census year	**Number of census records for linking**
1981	3143307
1986	3263284
1991	3373927
1996	3516513
2001	3630534

Table 2.3 Variables on the census linking files 1981–2001

Variable	**Type**	**Len**	**Format**	**Label**
AU	Char	6		Area Unit (Usual Residence Base 2001)
AgeC	Num	3	FAGE.	Age at census (years) – used for clerical review
Asian	Num	8	ETHA.	Ethnicity - Any Asian
AsianFix	Num	8	FIXA.	Asian - Ethnicity/Country of Birth Adjustment
BirthGp	Num	3	FBTHGP.	Country of Birth
DayC	Num	8		Day of Birth
ImpAge	Char	1	$FIMPAGE.	Age Imputation Indicator – used for clerical review
ImpSex	Num	8	FIMPSEX.	Sex Imputation Indicator – used for clerical review
MB	Char	7		Meshblock (Usual Residence Base 2001)
Maori	Num	8	ETHM.	Ethnicity - Any Maori i.e. Total Ethnicity
MonthC	Num	8	FMTH.	Month of Birth
PacFix	Num	8	FIXP.	Pacific - Ethnicity/Country of Birth Adjustment
Pacific	Num	8	ETHP.	Ethnicity – Any Pacific
Person_Id	Num	8 or 14		Census Person Id (length of 8 in 1996 and 2001, length of 14 in other years) – used for reference
YearC	Num	4		Year of Birth
nonMPA	Num	8	ETHO.	Ethnicity - Any NonMPA (European/Other)
sex	Num	8	FSEX.	Sex

2.4 Contents of cancer file

The cancer files were compiled from information from several different sources. The different sources may have different values for some attributes (for example, slightly different dates of birth). For this reason, some attributes have several variables, each with different information, with the information that is most likely to be correct recorded in the first variable. For example, a person with two dates of birth recorded as 04 October 1951 and 06 October 1951, with the first most likely to be correct, will have a value of 04 in variable Day1W and 06 in variable Day2W.

Table 2.4 Variables on the cancer linking files 1981–2001

Variable	Type	Len	Format	Label
AsianFixW	Num	3	FIXA.	Asian Ethnicity born in Asian from WSM data
AsianW	Num	3	ETHA.	Ethnicity - Any Asian on Cancer, Mort, NHI, Archived NHI files
BirthGpW	Num	3	FBTHGP.	Country of Birth Grouping from WSM
BirthGpW2	Num	3	FBTHGP.	Second Country of Birth Grouping from WSM – used for clerical review
Day1W	Num	3		Day of Birth Option 1 from WSM : Cancer and related data
Day2W	Num	3		Day of Birth Option 2 from WSM : Cancer and related data
Day3W	Num	3		Day of Birth Option 3 from WSM : Cancer and related data
Day4W	Num	3		Day of Birth Option 4 from WSM : Cancer and related data
Day5W	Num	3		Day of Birth Option 5 from WSM : Cancer and related data
MaoriW	Num	3	ETHM.	Ethnicity - Any Maori on Cancer, Mort, NHI, Archived NHI files
Month1W	Num	3	FMTH.	Month of Birth Option 1 from WSM : Cancer and related data
Month2W	Num	3	FMTH.	Month of Birth Option 2 from WSM : Cancer and related data
Month3W	Num	3	FMTH.	Month of Birth Option 3 from WSM : Cancer and related data
Month4W	Num	3	FMTH.	Month of Birth Option 4 from WSM : Cancer and related data
Month5W	Num	3	FMTH.	Month of Birth Option 5 from WSM : Cancer and related data
NumBthDates	Num	3		Number of Birth Dates on WSM data – used for clerical review
NumMbs	Num	3		Number of Meshblocks on Cancer Data various sources – used for clerical review
NumSexonWSM	Num	3		Number of Sex codes on WSM data – used for clerical review
PacFixW	Num	3	FIXP.	Pacific Ethnicity born in Pacific from WSM data
PacificW	Num	3	ETHP.	Ethnicity - Any Pacific on Cancer, Mort, NHI, Archived NHI files
PossDel	Num	3		If 1 then Possibly Delete as no address and suspect not New Zealand resident
SexW	Num	3	FSEX.	Sex from WSM : Cancer and related data
Year1W	Num	4		Year of Birth Option 1 from WSM : Cancer and related data
Year2W	Num	4		Year of Birth Option 2 from WSM : Cancer and related data
Year3W	Num	4		Year of Birth Option 3 from WSM : Cancer and related data
Year4W	Num	4		Year of Birth Option 4 from WSM : Cancer and related data
Year5W	Num	4		Year of Birth Option 5 from WSM : Cancer and related data
au1w	Char	6		Area Unit Option 1 from WSM : Cancer and related data Base 2001

au2w	Char	6		Area Unit Option 2 from WSM : Cancer and related data Base 2001
au3w	Char	6		Area Unit Option 3 from WSM : Cancer and related data Base 2001
au4w	Char	6		Area Unit Option 4 from WSM : Cancer and related data Base 2001
au5w	Char	6		Area Unit Option 5 from WSM : Cancer and related data Base 2001
au6w	Char	6		Area Unit Option 6 from WSM : Cancer and related data Base 2001
au7w	Char	6		Area Unit Option 7 from WSM : Cancer and related data Base 2011
au8w	Char	6		Area Unit Option 8 from WSM : Cancer and related data Base 2001
au9w	Char	6		Area Unit Option 9 from WSM : Cancer and related data Base 2001
cohort	Num	3		Cohort of Cancer Trends – used to keep track of dataset currently being used, and used for clerical review
id_num	Char	8		Cancer Trends Person Id
mb1w	Char	7		Meshblock Option 1 from WSM : Cancer and related data Base 2001
mb2w	Char	7		Meshblock Option 1 from WSM : Cancer and related data Base 2001
mb3w	Char	7		Meshblock Option 1 from WSM : Cancer and related data Base 2001
mb4w	Char	7		Meshblock Option 1 from WSM : Cancer and related data Base 2001
nonMPAW	Num	3	ETHO.	Ethnicity - Any nonMPA on Cancer, Mort, NHI, Archived NHI files
numBthGps	Num	3		Number of Countries of Birth on Cancer and related data – used for clerical review
numtaus	Num	3		Total Number of Area Units on Cancer Data various sources – used for clerical review

Notes:

(1) WSM = Wellington School of Medicine (source of information).
(2) Cancer = New Zealand Cancer Registry held by NZHIS (source of information).
(3) Mort = Mortality collection held by NZHIS (source of information).
(4) NHI = National Health Index (source of information).

The formats of the variables on the cancer files are the same as the corresponding variables on the census files.

The number of records for each year is given in the following list.

Table 2.5 Number of records on the cancer linking files 1981–2001

Census year	Number of cancer records	Number marked for deletion before linking (see PossDel variable)	Number left for linking
1981	52932	233	52699
1986	63856	230	63626
1991	77365	206	77159
1996	96583	141	96442
2001	83875	86	83789

3 Jobs and passes

A pass is the process used by QualityStage to link two files with a given specification of blocking variables, matching variables, m and u probabilities, and cut-off weight[2]. A job is a series of passes carried out, in a set order, on two initial files, producing one linked file. A job also produces two duplicate files and two residual files.

Duplicate links occur when one record has two or more links with records on the other file. QualityStage will place the record with the greatest weight in the linked file, and the others in the appropriate duplicate file. When the records have the same weight, QualityStage randomly selects one of the records for the linked file3. Census duplicates occur when one cancer record links with more than one census record. Cancer duplicates occur when one census record links with more than one cancer record.

Residual files are the remaining unlinked census and cancer records.

Several jobs were run, each refining the process, before the final linking job was run. Each job consists of several passes of the data, with unlinked residuals from each pass feeding into the next pass in the job. There may be two types of passes in each job, meshblock passes, and area unit passes. The features that are common to each type of pass are described here. (Each job and pass is documented in subsequent sections. The details that change with each job and pass will be included in that documentation.)

Two different types of jobs were run. The first type involved an iterative process of refining which variables should be used for blocking and which cut-off values should be used. At the end of this process, a final job was run from which all links were accepted for inclusion in the final linked file. After each job had been run decisions on the next job were made

The second type of job was run with the residuals of the final linking job. These jobs generated various links which were then considered for inclusion in the final linked file. This process is known as clerical review.

When multiple passes are used in one job, it is usually best practice to include the most effective pass first. Meshblock passes are more reliable than area unit passes because meshblock passes are less likely to produce false positive links than area unit passes. All of the jobs used in this project, where both meshblock and area unit passes were used, placed the meshblock passes before the area unit passes.

The values of the *m* and *u* probabilities were based on those used in the pilot study with 1996 data (Smith 2005).

Reports produced by QualityStage for each job and pass are not included in this document, but are available separately and are listed for each cohort. These reports assisted with the decisions on the final job from which all links were included in the final file.

[2] These and other technical terms are defined in the glossary at the end of this document.

[3] Section 6 describes fully how the project dealt with duplicates with the same weight.

3.1 Meshblock passes

The first type of pass blocks on meshblock. In these passes the variable MB from the census file is blocked with one of the variables MB1W, MB2W, or MB3W from the cancer file. The variables that are used for matching, with the MPROBs and UPROBs, are given in the table below.

Table 3.1 Variables used for matching in meshblock passes

Type of comparison			Variable A	Variable B	Mprob	Uprob
MATCH		CHAR	SEX	SEXW	0.99	0.5
MATCH	ARRAY	CHAR	DAYCA	DAY5A	0.97	0.033
MATCH	ARRAY	CHAR	MONTHCA	MONTH5A	0.98	0.083
MATCH	ARRAY	CHAR	YEARCA	YEAR5A	0.99	0.013
MATCH		CHAR	BIRTHGP	BRTHGPW	0.85	0.4
MATCH		CHAR	MAORI	MAORIW	0.8	0.5
MATCH		CHAR	PACIFIC	PACIFIW	0.8	0.5
MATCH		CHAR	ASIAN	ASIANW	0.8	0.5
MATCH		CHAR	NONMPA	NONMPAW	0.8	0.5
MATCH		CHAR	PACFIX	PACFIXW	0.92	0.033
MATCH		CHAR	ASNFIX	ASNFIXW	0.92	0.033

The match variables were processed in the order listed in the above table. (Note that the links made depend on the order that the match variables are processed in.)

3.2 Area unit passes

The second type of pass blocks on area unit. In these passes the variable AU from the census file is blocked with one of the variables AU1W, AU2W, AU3W or AU4W from the cancer file. The variables that are used for matching, with the MPROBs and UPROBs, are given in the table below.

The population of area units is too large for area unit to be used by QualityStage as a blocking variable on its own. The variables SEX from the census file and SEXW from the cancer file are also used as blocking variables in these passes. Using sex as a second blocking variable effectively halves the size of the populations compared to aid QualityStage to make the links. Blocking variables are not compared probabilistically so sex was chosen as a second blocking variable because it is the most reliable matching variable.

Table 3.2 Variables used for matching in area unit passes

Type of comparison			Variable A	Variable B	Mprob	Uprob
MATCH	ARRAY	CHAR	DAYCA	DAY5A	0.97	0.033
MATCH	ARRAY	CHAR	MONTHCA	MONTH5A	0.98	0.083
MATCH	ARRAY	CHAR	YEARCA	YEAR5A	0.99	0.013
MATCH		CHAR	BIRTHGP	BRTHGPW	0.85	0.4
MATCH		CHAR	MAORI	MAORIW	0.8	0.5
MATCH		CHAR	PACIFIC	PACIFIW	0.8	0.5
MATCH		CHAR	ASIAN	ASIANW	0.8	0.5
MATCH		CHAR	NONMPA	NONMPAW	0.8	0.5
MATCH		CHAR	PACFIX	PACFIXW	0.92	0.033
MATCH		CHAR	ASNFIX	ASNFIXW	0.92	0.033

The match variables were processed in the order listed in the above table.

4 Positive predictive value (PPV)

The purpose of this project was to find census records that matched cancer records. Because this project used probabilistic linking, it is not possible to be certain that all links are matches. (A match is a pair of records that apply to the same individual while a link is a pair of records that are highly likely to apply to the same individual.) The final linked file will therefore contain false positive links and will not contain false negative links (see the following diagram).

Figure 4.1 Links and matches

	Matches	Non-matches
Linked	True positives	False positives
Unlinked	False negatives	True negatives

Minimising the number of false positive links improves the quality and usefulness of the linked file (Blakely and Salmond 2002 p 1248). The positive predictive value (PPV) was used to estimate the number of false positives for passes, assisting decisions on final jobs and passes.

The (PPV) is defined as the proportion of linked records that are true positives. A high PPV indicates that the proportion of false positive links is small. The PPV of each pass in QualityStage was calculated using a method described in Blakely and Salmond 2002.

At this stage it has not been possible to calculate PPVs for the links added in the clerical review. The clerical review was done in SAS, not QualityStage, and the Blakely and Salmond calculation uses information from the QualityStage passes. It might have been possible to conduct the clerical review using QualityStage alone, and therefore to calculate PPVs, however, the strengths of the clerical review method used compared to any method developed in QualityStage outweigh the unavailability of the PPVs.

The PPV values for each of the final passes are shown in the tables summarising the links for each year. Reports showing detailed results of PPV calculations for each pass (for interim and final jobs) are not included in this document, but are available separately and are listed for each cohort. These PPV reports assisted with making the decisions on the final job used to produce links included in the final file.

5 The clerical review process

5.1 Overview of the process

All of the links from the clerical review jobs were read into SAS datasets and concatenated into one SAS dataset. Records with the same weight as a duplicate record were removed[4]. A variable that summarised the values of variables used in linking was created for each linked record (COMB variable). Tables of the COMB variable and various weights were produced. Using these tables, uniform criteria for which records should be included or excluded were developed. The clerical review file of good links was then added to the link file with good links identified in the final job in QualityStage.

5.2 SAS dataset

The SAS dataset contained all of the census variables, followed by all of the cancer variables, and extra variables related to the matching process. These extra variables were:

From QualityStage

- ID ID number from the job (duplicates and the corresponding links have the same ID).
- Wgt weight.

Added in SAS

- Pass Pass where match was made (passes 1 to 3 from job 4, passes 4 to 8 from job 5, passes 9 and 10 from job 6, and passes 11 to 12 from job 7).
- MBAU Indicator of mesh block or area unit pass.
- COMB Composite variable based on flags indicating if matching variables agreed, disagreed, or were missing.
- YEARTOL Smallest difference between year of birth on census record and any year of birth on cancer record.

5.3 COMB variable

The COMB variable consists of 13 columns, one for each indicator variable. Each column will contain a letter if the values of the variable from each source record agreed on the linked record, an x if they disagreed, and "." if the values were missing.

The "tolerance in year" column is slightly different to the others as it compares the year of birth from the census and cancer records, but records a value of "T" in column 5 if the value of the year of birth on the census record is the same or one year different from any of the years of birth on the cancer record. For example, census records with a year of birth of 1950 linked to cancer records with any year of birth recorded as 1949, 1950, and 1951 would all have "T" recorded in column 5 of the COMB variable. All other values of year of birth in the cancer records would have "X" recorded in column 5, or "." if the year of birth was missing from the cancer record. This column indicates that different dates of birth may be recorded for the

[4] Section 6 describes fully how the project dealt with duplicates with the same weight.

same person, and that a difference of one year is acceptable in some circumstances. (See the "Criteria used to include or exclude clerically reviewed records" sections for each cohort for details of how this column of the COMB variable is used.)

Table 5.1 Variables represented in each column of the COMB variable

Column	Variable	Letter used	Notes
1	Sex	S	
2	Day	D	Day from date of birth
3	Month	M	Month form date of birth
4	Year	Y	Year from date of birth
5	Tolerance in year	T	Indicator if year of birth is up to one year different
6	Country of birth	C	
7	Summary of ethnicity	E	
8	Maori	m	
9	Pacific	p	
10	Asian	a	
11	Non MPA	n	
12	Pacific fix	P	
13	Asian fix	A	

The COMB variable, in combination with the weight variable, assists with assessing the quality of a match as it is possible to see at a glance which variables matched.

Various tables of the COMB variable and weights were produced (as an example, part of a table is produced below showing weights between 12 and 13 rounded to one decimal place).

Table 5.2 Example of a table using the COMB variable

```
„ƒƒƒƒƒƒƒƒƒƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ†
,              ,     ,                             wgtrnd                                ,
,              ,     ‡ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ…ƒƒƒƒƒ‰
,              , All , 12  ,12.1 ,12.2 ,12.3 ,12.4 ,12.5 ,12.6 ,12.7 ,12.8 ,12.9 , 13  ,
‡ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒˆƒƒƒƒƒ‰
,comb          ,     ,     ,     ,     ,     ,     ,     ,     ,     ,     ,     ,     ,
,S..YT.Empan.. ,    3,    .,    .,    1,    .,    .,    1,    1,    .,    .,    .,    .,
,S..YT.Empan.A ,    1,    .,    .,    .,    .,    .,    .,    .,    1,    .,    .,    .,
,S..YTCEmpan.. ,    2,    .,    .,    .,    .,    .,    1,    1,    .,    .,    .,    .,
,S.MYT.Empan.. ,    8,    .,    2,    1,    .,    1,    .,    .,    .,    3,    .,    1,
,S.MYTCEmpan.. ,    3,    .,    .,    .,    .,    .,    2,    .,    .,    .,    .,    1,
,SD.YT.Empan.. ,    6,    .,    2,    1,    .,    .,    2,    .,    .,    1,    .,    .,
,SD.YT.Emxan.. ,    1,    .,    .,    .,    .,    .,    .,    .,    .,    .,    .,    1,
,SD.YTCEmpan.. ,    1,    .,    1,    .,    .,    .,    .,    .,    .,    .,    .,    .,
Šƒƒƒƒƒƒƒƒƒƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒ‹ƒƒƒƒƒŒ
```

5.4 YEARTOL variable

The YEARTOL (year tolerance) variable measures the difference between the year of birth on the census record and the cancer record. Because the cancer records may have multiple years of birth recorded, the YEARTOL is defined as the smallest difference between the year of birth on the census record and any year of birth on the cancer record.

Year tolerance also forms part of the COMB variable defined above where a difference of one year is counted as an agreement. Both of these measures were used in the clerical review process.

5.5 Criteria used in clerical review decisions

The tables with the COMB variable assisted in creating criteria used to include records representing good links.

For each cohort, similar uniform criteria were developed. The approach of using these criteria, applied to each record, rather than examining each record individually, ensured that decisions were consistent.

6 Resolution of duplicate pairs and identical links

6.1 Duplicate pairs with the same weight

Sometimes QualityStage links one cancer record to multiple census records, or one census record to multiple cancer records, and the links have the same weight. In these situations, one of the records is randomly chosen as a link, and the others are classified as duplicates. Links that had duplicates with the same weight were identified and removed from the link file. (The reason for removing these records is that there is at best a 50% chance that the chosen link is correct.)

These duplicates were removed at two stages. First, after job 3 produced good links for the final file. Secondly, after jobs 4 to 7 produced records for clerical review.

Table 6.1 Numbers of duplicate links removed 1981–2001

Year	Duplicates removed
1981	229
1986	244
1991	382
1996	474
2001	308

6.2 Identical links

Identical links occur when the same cancer and census record are linked in different QualityStage jobs. Identical links may have occurred because of the process used for clerical review. Several QualityStage jobs were run with the same input files and the links from all of these jobs were clerically reviewed. It is possible that the same records linked in different jobs. When the records from job 3 (the final QualityStage job) and the records accepted as links after clerical review were joined, records with the same id number (cancer and or census) were removed.

2001 Cohort

2001.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 24 July 2007

Table 2001.1.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 2001.1.2 Summary of match

Pass number and description	**In**		**out**	
1 (block on MB1)	A CTCensus	3630534	MATCH	64621
	B Cancer	83789	DUPA	35644
			DUPB	949
			RESA	3530269
			RESB	18219
2 (block on MB2)	RESA	3530269	MATCH	4609
	RESB	18219	DUPA	2020
			DUPB	22
			RESA	3523640
			RESB	13588
3 (Block on MB3)	RESA	3523640	MATCH	511
	RESB	13588	DUPA	203
			DUPB	5
			RESA	3522926
			RESB	13072
Summary	A CTCensus	3630534	LINK1	69741
	B Cancer	83789	DUPA	37867
			DUPB	976
			RESA	3522926
			RESB	13072

Note DUP = duplicates and RES = residuals

2001.1.1 Decisions

These decisions were based on the QualityStage and PPV reports, not included in this documentation, but listed in sections 2001.10 and 2001.11.

1 Pass three, blocking on MB3W from the cancer file should not be used because there were only a small number of links (511) and the weights were too low. Good links can be picked up in clerical review.

2 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 12.99.

3 Four area unit passes, with no cut-off value should be run.

2001.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 25 July 2007

Table 2001.2.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	12.99
2	MB	MB2W	12.99
3	AU and SEX	AU1W and SEXW	0
4	AU and SEX	AU2W and SEXW	0
5	AU and SEX	AU3W and SEXW	0
6	AU and SEX	AU4W and SEXW	0

Table 2001.2.2 Summary of match

Pass number and description	**In**		**Out**	
1 (block on MB1)	A CTCensus	3630534	MATCH	50132
	B Cancer	83789	DUPA	169
			DUPB	24
			RESA	3580233
			RESB	33633
2 (block on MB2)	RESA	3580233	MATCH	5160
	RESB	33633	DUPA	18
			DUPB	0
			RESA	3575055
			RESB	28473
3 (block on AU1)	RESA	3575055	MATCH	27845
	RESB	28473	DUPA	174536
			DUPB	245
			RESA	3372674
			RESB	383
4 (block on AU2)	RESA	3372674	MATCH	322
	RESB	383	DUPA	1483
			DUPB	0
			RESA	3370869
			RESB	61
5 (block on AU3)	RESA	3370869	MATCH	18
	RESB	61	DUPA	73
			DUPB	0
			RESA	3370778
			RESB	43

6 (block on AU4)	RESA	3370778	MATCH	4
	RESB	43	DUPA	15
			DUPB	0
			RESA	3370759
			RESB	39
Summary	A CTCensus	3630534	LINK1	83481
	B Cancer	83789	DUPA	176294
			DUPB	269
			RESA	3370759
			RESB	39

2001.2.1 Decisions

1 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 12.99.

2 Passes 3 and 4 (blocking on AU1W and AU2W) should be rerun, with a cut-off value of 14.60.

3 Passes 5 and 6 (blocking on AU3W and AU4W) produce only a small number of links (18 and four respectively) and should not be used.

2001.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 27 July 2007

Table 2001.3.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	12.99
2	MB	MB2W	12.99
3	AU and SEX	AU1W and SEXW	14.60
4	AU and SEX	AU2W and SEXW	14.60

Table 2001.3.2 Summary of match

Pass number and description	**In**		**Out**	
1 (block on MB1)	A CTCensus	3630534	MATCH	50132
	B Cancer	83789	DUPA	169
			DUPB	24
			RESA	3580233
			RESB	33633
2 (block on MB2)	RESA	3580233	MATCH	5160
	RESB	33633	DUPA	18
			DUPB	0
			RESA	3575055
			RESB	28473
3 (block on AU1)	RESA	3575055	MATCH	4720
	RESB	28473	DUPA	197
			DUPB	3
			RESA	3570138
			RESB	23750
4 (block on AU2)	RESA	3570138	MATCH	2852
	RESB	23750	DUPA	118
			DUPB	1
			RESA	3567168
			RESB	20897
Summary	A CTCensus	3630534	LINK1	62864
	B Cancer	83789	DUPA	502
			DUPB	28
			RESA	3567168
			RESB	20897

2001.3.1 Decisions

1 Links from this job will be included in the final linked file. (62864 linked files, or 75.03% of cancer files.)

2 Use residuals from this job (that is 3567168 census residuals and 20897 cancer residuals) for four jobs to produce records for clerical review. These jobs were considered the most likely to produce some good links that could be included in the final linked file.

3 Job 4, three passes blocking on MB1 to MB3 with no cut-off.

4 Job 5, four passes blocking on AU1 to AU4 with no cut-off.

5 Job 6, two passes blocking on MB2 and AU2 with no cut-off.

6 Job 7, two passes blocking on MB3 and AU3 with no cut-off.

2001.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 1 August 2007

Table 2001.4.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 2001.4.2 Summary of match

Pass number and description	**In**		**Out**	
1 (block on MB1)	RESA1 (Census)	3567168	MATCH	9998
	RESB1 (Cancer)	20897	DUPA	6700
			DUPB	347
			RESA	3550470
			RESB	10552
2 (block on MB2)	RESA	3550470	MATCH	1679
	RESB	10552	DUPA	1001
			DUPB	17
			RESA	3547790
			RESB	8856
3 (block on MB3)	RESA	3547790	MATCH	329
	RESB	8856	DUPA	175
			DUPB	1
			RESA	3547286
			RESB	8526
Summary	RESA1 (Census)	3567168	LINK1	12006
	RESB1 (Cancer)	20897	DUPA	7876
			DUPB	365
			RESA	3547286
			RESB	8526

`

2001.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 1 August 2007

Table 2001.5.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	AU and SEX	AU1W and SEXW	0
2	AU and SEX	AU2W and SEXW	0
3	AU and SEX	AU3W and SEXW	0
4	AU and SEX	AU4W and SEXW	0

Table 2001.5.2 Summary of match

Pass number and description	**In**		**Out**	
1 (block on AU1)	RESA1 (Census)	3567168	MATCH	20398
	RESB1 (Cancer)	20897	DUPA	133225
			DUPB	141
			RESA	3413545
			RESB	358
2 (block on AU2)	RESA	3413545	MATCH	279
	RESB	358	DUPA	1285
			DUPB	0
			RESA	3411981
			RESB	79
3 (block on AU3)	RESA	3411981	MATCH	21
	RESB	79	DUPA	87
			DUPB	0
			RESA	3411873
			RESB	58
4 (block on AU4)	RESA	3411873	MATCH	2
	RESB	58	DUPA	9
			DUPB	0
			RESA	3411862
			RESB	56
Summary	RESA1 (Census)	3567168	LINK1	20700
	RESB1 (Cancer)	20897	DUPA	134606
			DUPB	141
			RESA	3411862
			RESB	56

2001.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 10 August 2007

Table 2001.6.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB2W	0
2	AU and SEX	AU2W and SEXW	0

Table 2001.6.2 Summary of match

Pass number and description	**In**		**Out**	
1 (block on MB2)	RESA1 (Census)	3567168	MATCH	4298
	RESB1 (Cancer)	20897	DUPA	3286
			DUPB	120
			RESA	3559584
			RESB	16479
2 (block on AU2)	RESA	3559584	MATCH	10300
	RESB	16479	DUPA	71077
			DUPB	38
			RESA	3478207
			RESB	6141
Summary	RESA1 (Census)	3567168	LINK1	14598
	RESB1 (Cancer)	20897	DUPA	74363
			DUPB	158
			RESA	3478207
			RESB	6141

2001.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 14 September 2007

Table 2001.7.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB3W	0
2	AU and SEX	AU3W and SEXW	0

Table 2001.7.2 Summary of match

Pass number and description	**In**		**Out**	
1 (block on MB3)	RESA1 (Census)	3567168	MATCH	1238
	RESB1 (Cancer)	20897	DUPA	930
			DUPB	14
			RESA	3565000
			RESB	19645
2 (block on AU3)	RESA	3565000	MATCH	6919
	RESB	19645	DUPA	56874
			DUPB	15
			RESA	3501207
			RESB	12711
Summary	RESA1 (Census)	3567168	LINK1	8157
	RESB1 (Cancer)	20897	DUPA	57804
			DUPB	29
			RESA	3501207
			RESB	12711

2001.8 Clerical review process

2001.8.1 Criteria used to include or exclude clerically reviewed records

Applying the criteria to decide if a clerically reviewed record should be included or excluded was an iterative process and in some cases resulted in attaching an interim flag to some records. The values of the interim flag were used to decide if a record should be finally included or excluded when later criteria were applied. A final flag was then attached to each record to indicate if it should be included or excluded from the final file.

Values of interim flag:

- Not a link
- Investigate possible link
- Probably not a link
- Look at - High Weight
- Make a link - Investigated & acceptable

In these criteria, records given an interim or final flag of "Not a link" are described as deleted. Records given an interim or final flag value of "Make a link - Investigated & acceptable" are described with the phrase "make a link".

Values of final flag:

- Not a link
- Make a link - Investigated & acceptable

Many of these criteria use parts of the COMB variable (eg day or month). A full description of the COMB variable is given in section 5.3.

In these criteria the phrase "year tolerance" is used in two different ways. First, as part of the COMB variable "year tolerance agrees" means that there is no difference in the year of birth between the two linked records or the difference is one year. Similarly, "Year tolerance disagrees" means that the difference is more than one year. Secondly, the YEARTOL variable measures the actual difference in years and the use of this variable is indicated in phrases like "year tolerance greater then 5". See sections 5.3 and 5.4 for fuller descriptions of the COMB and YEARTOL variables.

Criteria used to include or exclude clerically reviewed records:

1 Records with a weight below 5 were deleted.

2 Records linked in an AU pass with a weight below 8.5 were deleted.

3 Duplicate records with the same weight were deleted.

4 Records with a weight of 10 or more were flagged as "Look at - High Weight".

5 For records linked in an MB pass, with weights between 5 and 7.99, the following criteria were applied (in order):

- If year and year tolerance part of COMB variable indicate missing values then delete.
- If value of year tolerance variable greater than 5 then delete.
- If ethnicity values (Empan columns of COMB variable) disagree then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagree or year disagree or country of birth disagree then delete.
- else flag as "Probably not a link".

6 For records linked in an AU pass, with weights between 8.5 and 8.99, following criteria were applied (in order):

- If year and year tolerance missing then delete.
- If year tolerance greater than 1 then delete.
- If ethnicity (Empan) disagrees then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagrees or year disagrees, or country of birth disagrees then delete.
- else flag as "Probably not a link".

7 For records linked in an MB pass, the following criteria were applied (in order):

- For weights greater than or equal to 10 where day and month agree, or year or year tolerance agree, make a link.
- For weights with interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" and month, year and ethnicity agree, make a link.
- For weights greater than or equal to 10 and date of birth agrees (day, month and year) or day, month and year tolerance, or month and year agree, or day and year agree, or day and month and ethnicity agree, then make a link.

8 For records linked in an MB pass, where ethnicity agrees, and interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" the following criteria were applied (in order):

- If month and year agree, make a link.
- If day, month and year tolerance agree, make a link.
- If day and year agree, make a link.

9 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and ethnicity agrees or missing, make a link.

10 Records with an interim flag value missing, or a value of "Probably not a link" and date missing were deleted.

11 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight" and weights greater than or equal to 7, and sex and ethnicity agree, the following criteria were applied (in order):

- If country missing or agrees and, either, but not both, day or month disagree or missing, make a link.

- If pacific fix or asian fix agree, and either day and/or country disagree or missing, make a link.

12 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Look at - High Weight", the following criteria were applied (in order):

- If ethnicity agrees for Maori, Pacific and Asian (ignore nonMPA column in COMB variable) and year agrees, or day month and year tolerance agree, then make a link.

13 For records linked in an MB pass, with a weight greater than or equal to 7.99 and with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", the following criteria were applied (in order):

- If year is missing, but year tolerance is less than 5, and sex, day, month, country, Maori, Pacific, and Asian agree, then make a link.
- If date of birth all missing, but year tolerance less than 2 and sex, country and ethnicity agree, then make a link.
- If sex and year are agree, but ethnicity missing and country missing or agrees, then make a link.

14 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and year and year tolerance disagree (and country agrees or missing) the following criteria were applied (in order):

- If records linked in an MB pass and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and year tolerance is less than 2, then make a link.

15 For records with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Probably not a link" or "Look at - High Weight", and year missing and ethnicity (other than NonMPA part of COMB variable) agree or missing, the following criteria were applied (in order):

- If records linked in an MB pass and weight greater than 7.99 and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and weight greater than 8.99 and year tolerance is less than 3, then make a link.

16 For records with interim flag value "Look at - High Weight", and weight greater than or equal to 10, and year tolerance less than 2, then make a link.

17 For records with interim flag value missing, or values of "Strong possible link", or " Investigate possible link" or "Probably not a link" or "Look at - High Weight", the following criteria were applied (in order):

- If weight less than or equal to 7.99 then delete.
- If records linked in an MB pass and year tolerance is greater than 5, then delete.
- If records linked in an AU pass and (weight greater than 8.99 or year tolerance is greater than 1), then delete.

18 For the remaining records, assign the value of the final flag:

- For records with interim flag value of "Investigate possible link", then make a link.
- For records with interim flag value missing, or values of "Probably not a link" or "Look at - High Weight", then delete.

Table 2001.8.1 Summary of results of clerical review

Type of pass	Number linked in pass	Cumulative number linked (including final job)	% of cancer files
MB	2994	65858	78.60%
AU	2884	68742	82.04%

2001.9 Summary of links

Table 2001.9.1 Summary of links for 2001

Pass	Pass details	Number linked in pass	Cumulative number linked	Cumulative % of cancer files	PPV
1	MB1	50132	50132	59.83%	99.8
2	MB2	5160	55292	65.99%	99.2
3	AU1, sex	4720	60012	71.62%	75.9
4	AU2, sex	2852	62864	75.03%	75.4
CR MB	MB1 – MB3, MB2 (with AU2), MB3 (with AU3)	2994	65858	78.60%	
CR AU	AU1 – AU4, AU2 (after MB2), AU3 (after MB3)	2884	68742	82.04%	
Dups	Delete duplicates with identical weights	(308)	68434	81.67%	

CR MB – Clerical review records added in meshblock passes.
CR AU – Clerical review records added in area unit passes.
PPVs are not available for the clerical review passes.

2001.10 List of QualityStage reports for each job and pass

2001.10.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 24 July 2007

MatchOn123MB_2001_1.out
MatchOn123MB_2001_2.out
MatchOn123MB_2001_3.out

2001.10.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 25 July 2007

MatchOn12MBwco12pt99_3to6AUnco_2001_1.out
MatchOn12MBwco12pt99_3to6AUnco_2001_2.out
MatchOn12MBwco12pt99_3to6AUnco_2001_3.out
MatchOn12MBwco12pt99_3to6AUnco_2001_4.out
MatchOn12MBwco12pt99_3to6AUnco_2001_5.out
MatchOn12MBwco12pt99_3to6AUnco_2001_6.out

2001.10.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 27 July 2007

MatchOn12MBwco12pt99_34AUwco14pt60_2001_1.out
MatchOn12MBwco12pt99_34AUwco14pt60_2001_2.out
MatchOn12MBwco12pt99_34AUwco14pt60_2001_3.out
MatchOn12MBwco12pt99_34AUwco14pt60_2001_4.out

2001.10.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 1 August 2007

MatchOnRESp4MB123NOCU_2001_1.out
MatchOnRESp4MB123NOCU_2001_2.out
MatchOnRESp4MB123NOCU_2001_3.out

2001.10.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 1 August 2007

MatchOnRESp4AU1234NOCU_2001_1.out
MatchOnRESp4AU1234NOCU_2001_2.out
MatchOnRESp4AU1234NOCU_2001_3.out
MatchOnRESp4AU1234NOCU_2001_4.out

2001.10.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 10 August 2007

MatchOnRESp4MB2AU2NOCO_2001_1.out
MatchOnRESp4MB2AU2NOCO_2001_2.out

2001.10.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 14 September 2007

MatchOnRESp4MB3AU3NOCO_2001_1.out
MatchOnRESp4MB3AU3NOCO_2001_2.out

2001.11 List of PPV calculations for each job and pass

2001.11.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 24 July 2007

Passes123MB NCO_2001_lastppvPass1.out
Passes123MB NCO_2001_lastppvPass2.out
Passes123MB NCO_2001_lastppvPass3.out

2001.11.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 25 July 2007

Passes12MBwco12pt99_3to6AUnco_2001_lastppvPass1.out
Passes12MBwco12pt99_3to6AUnco_2001_lastppvPass2.out
Passes12MBwco12pt99_3to6AUnco_2001_lastppvPass3.out
Passes12MBwco12pt99_3to6AUnco_2001_lastppvPass4.out
Passes12MBwco12pt99_3to6AUnco_2001_lastppvPass5.out
Passes12MBwco12pt99_3to6AUnco_2001_lastppvPass6.out

2001.11.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 27 July 2007

Passes12MBwco12pt99_34AUwco14pt60_2001_lastppvPass3.out
Passes12MBwco12pt99_34AUwco14pt60_2001_lastppvPass4.out

1996 Cohort

1996.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 24 July 2007

Table 1996.1.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1996.1.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3516513	MATCH	68038
	B Cancer	96442	DUPA	41720
			DUPB	1927
			RESA	3406755
			RESB	26477
2 Block on MB2	RESA	3406755	MATCH	2472
	RESB	26477	DUPA	1179
			DUPB	9
			RESA	3403104
			RESB	23996
3 Block on MB3	RESA	3403104	MATCH	138
	RESB	23996	DUPA	51
			DUPB	0
			RESA	3402915
			RESB	23858
Summary	A CTCensus	3516513	LINK1	70648
	B Cancer	96442	DUPA	42950
			DUPB	1936
			RESA	3402915
			RESB	23858

1996.1.1 Decisions

1 Pass three, blocking on MB3W from the cancer file should not be used because there were only a small number of links (138) and the weights were too low. Good links can be picked up in clerical review.

2 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 11.99.

3 Four area unit passes, with no cut-off value should be run.

1996.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 25 July 2007

Table 1996.2.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	11.99
2	MB	MB2W	11.99
3	AU and SEX	AU1W and SEXW	0
4	AU and SEX	AU2W and SEXW	0
5	AU and SEX	AU3W and SEXW	0
6	AU and SEX	AU4W and SEXW	0

Table 1996.2.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3516513	MATCH	47502
	B Cancer	96442	DUPA	218
			DUPB	22
			RESA	3468793
			RESB	48918
2 Block on MB2	RESA	3468793	MATCH	2102
	RESB	48918	DUPA	15
			DUPB	1
			RESA	3466676
			RESB	46815
3 Block on AU1	RESA	3466676	MATCH	45466
	RESB	46815	DUPA	265574
			DUPB	589
			RESA	3155636
			RESB	760
4 Block on AU2	RESA	3155636	MATCH	589
	RESB	760	DUPA	2186
			DUPB	3
			RESA	3152861
			RESB	168
5 Block on AU3	RESA	3152861	MATCH	94
	RESB	168	DUPA	374
			DUPB	3
			RESA	3152393
			RESB	71
6 Block on AU4	RESA	3152393	MATCH	2
	RESB	71	DUPA	0
			DUPB	0
			RESA	3152391
			RESB	69

Summary	A CTCensus	3516513	LINK1	95755
	B Cancer	96442	DUPA	268367
			DUPB	618
			RESA	3152391
			RESB	69

1996.2.1 Decisions

1 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 11.99.

2 Passes 3 and 4 (blocking on AU1W and AU2W) should be rerun, with a cut-off value of 14.60.

3 Passes 5 and 6 (blocking on AU3W and AU4W on cancer file) should not be used because they produce only a small number of links (94 and 2) and the weights were too low. Good links will be picked up in clerical review.

1996.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 27 July 2007

Table 1996.3.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	11.99
2	MB	MB2W	11.99
3	AU and SEX	AU1W and SEXW	14.60
4	AU and SEX	AU2W and SEXW	14.60

Table 1996.3.2 Summary of match

Pass number and description	**In**		**out**	
1 Block on MB1	A CTCensus	3516513	MATCH	47502
	B Cancer	96442	DUPA	218
			DUPB	22
			RESA	3468793
			RESB	48918
2 Block on MB2	RESA	3468793	MATCH	2102
	RESB	48918	DUPA	15
			DUPB	1
			RESA	3466676
			RESB	46815
3 Block on AU1	RESA	3466676	MATCH	9058
	RESB	46815	DUPA	324
			DUPB	9
			RESA	3457294
			RESB	37748
4 Block on AU2	RESA	3457294	MATCH	5975
	RESB	37748	DUPA	228
			DUPB	5
			RESA	3451091
			RESB	31768
Summary	A CTCensus	3516513	LINK1	64637
	B Cancer	96442	DUPA	785
			DUPB	37
			RESA	3451091
			RESB	31768

1996.3.1 Decisions

1 Links from this job will be included in the final linked file. (64637 linked files, or 67.02% of cancer files.)

2 Use residuals from this job (that is 3451091 census residuals and 31768 cancer residuals) for four jobs to produce records for clerical review. These jobs were considered the most likely to produce some good links that could be included in the final linked file.

3 Job 4, three passes blocking on MB1 to MB3 with no cut-off.

4 Job 5, four passes blocking on AU1 to AU4 with no cut-off.

5 Job 6, two passes blocking on MB2 and AU2 with no cut-off.

6 Job 7, two passes blocking on MB3 and AU3 with no cut-off.

1996.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 1 August 2007

Table 1996.4.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1996.4.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	RESA1 (Census)	3451091	MATCH	14755
	RESB1 (Cancer)	31768	DUPA	12324
			DUPB	775
			RESA	3424012
			RESB	16238
2 Block on MB2	RESA	3424012	MATCH	1315
	RESB	16238	DUPA	903
			DUPB	6
			RESA	3421794
			RESB	14917
3 Block on MB3	RESA	3421794	MATCH	107
	RESB	14917	DUPA	88
			DUPB	0
			RESA	3421599
			RESB	14810
Summary	RESA1 (Census)	3451091	LINK1	16177
	RESB1 (Cancer)	31768	DUPA	13315
			DUPB	781
			RESA	3421599
			RESB	14810

1996.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 1 August 2007

Table 1996.5.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	AU and SEX	AU1W and SEXW	0
2	AU and SEX	AU2W and SEXW	0
3	AU and SEX	AU3W and SEXW	0
4	AU and SEX	AU4W and SEXW	0

Table 1996.5.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on AU1	RESA1 (Census)	3451091	MATCH	30696
	RESB1 (Cancer)	31768	DUPA	190053
			DUPB	378
			RESA	3230342
			RESB	694
2 Block on AU2	RESA	3230342	MATCH	512
	RESB	694	DUPA	1884
			DUPB	2
			RESA	3227946
			RESB	180
3 Block on AU3	RESA	3227946	MATCH	96
	RESB	180	DUPA	414
			DUPB	1
			RESA	3227436
			RESB	83
4 Block on AU4	RESA	3227436	MATCH	5
	RESB	83	DUPA	6
			DUPB	0
			RESA	3227425
			RESB	78
Summary	RESA1 (Census)	3451091	LINK1	31309
	RESB1 (Cancer)	31768	DUPA	192357
			DUPB	381
			RESA	3227425
			RESB	78

1996.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 13 August 2007

Table 1996.6.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB2W	0
2	AU and SEX	AU2W and SEXW	0

Table 1996.6.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB2	RESA1 (Census)	3451091	MATCH	3841
	RESB1 (Cancer)	31768	DUPA	3890
			DUPB	111
			RESA	3443360
			RESB	27816
2 Block on AU2	RESA	3443360	MATCH	18982
	RESB	27816	DUPA	124264
			DUPB	214
			RESA	3300114
			RESB	8620
Summary	RESA1 (Census)	3451091	LINK1	22823
	RESB1 (Cancer)	31768	DUPA	128154
			DUPB	325
			RESA	3300114
			RESB	8620

1996.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 14 September 2007

Table 1996.7.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB3W	0
2	AU and SEX	AU3W and SEXW	0

Table 1996.7.2 Summary of match

Pass number and description	In		Out	
1 Block on MB3	RESA1 (Census)	3451091	MATCH	496
	RESB1 (Cancer)	31768	DUPA	654
			DUPB	3
			RESA	3449941
			RESB	31269
2 Block on AU3	RESA	3449941	MATCH	12875
	RESB	31269	DUPA	100475
			DUPB	106
			RESA	3336591
			RESB	18288
Summary	RESA1 (Census)	3451091	LINK1	13371
	RESB1 (Cancer)	31768	DUPA	101129
			DUPB	109
			RESA	3336591
			RESB	18288

1996.8 Clerical review process

1996.8.1 Criteria used to include or exclude clerically reviewed records

Applying the criteria to decide if a clerically reviewed record should be included or excluded was an iterative process and in some cases resulted in attaching an interim flag to some records. The values of the interim flag were used to decide if a record should be finally included or excluded when later criteria were applied. A final flag was then attached to each record to indicate if it should be included or excluded from the final file.

Values of interim flag:

- Not a link
- Investigate possible link
- Probably not a link
- Look at - High Weight
- Make a link - Investigated & acceptable

In these criteria, records given an interim or final flag of "Not a link" are described as deleted. Records given an interim or final flag value of "Make a link - Investigated & acceptable" are described with the phrase "make a link".

Values of final flag:

- Not a link
- Make a link - Investigated & acceptable

Many of these criteria use parts of the COMB variable (eg day or month). A full description of the COMB variable is given in section 5.3.

In these criteria the phrase "year tolerance" is used in two different ways. First, as part of the COMB variable "year tolerance agrees" means that there is no difference in the year of birth between the two linked records or the difference is one year. Similarly, "Year tolerance disagrees" means that the difference is more than one year. Secondly, the YEARTOL variable measures the actual difference in years and the use of this variable is indicated in phrases like "year tolerance greater then 5". See sections 5.3 and 5.4 for fuller descriptions of the COMB and YEARTOL variables.

Criteria used to include or exclude clerically reviewed records:

1 Records with a weight below 4 were deleted.

2 Records linked in an AU pass with a weight below 4.3 were deleted.

3 Duplicate records with the same weight were deleted.

4 Records with a weight of 9 or more were flagged as "Look at - High Weight".

5 For records linked in an MB pass, with weights between 4 and 4.99, the following criteria were applied (in order):

- If year and year tolerance part of COMB variable indicate missing values then delete.
- If value of year tolerance variable greater than 5 then delete.
- If ethnicity values (Empan columns of COMB variable) disagree then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagree or year disagree or country of birth disagree then delete.
- else flag as "Probably not a link".

6 For records linked in an AU pass, with weights between 4.3 and 6.49, following criteria were applied (in order):

- If year and year tolerance missing then delete.
- If year tolerance greater than 1 then delete.
- If ethnicity (Empan) disagrees then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagrees or year disagrees, or country of birth disagrees then delete.
- else flag as "Probably not a link".

7 For records linked in an MB pass, the following criteria were applied (in order):

- For weights greater than or equal to 9 where day and month agree, or year or year tolerance agree, make a link.
- For weights with interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" and month, year and ethnicity agree, make a link.
- For weights greater than or equal to 9 and date of birth agrees (day, month and year) or day, month and year tolerance, or month and year agree, or day and year agree, or day and month and ethnicity agree, then make a link.

8 For records linked in an MB pass, where ethnicity agrees, and interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" the following criteria were applied (in order):

- If month and year agree, make a link.
- If day, month and year tolerance agree, make a link.
- If day and year agree, make a link.

9 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and ethnicity agrees or missing, make a link.

10 Records with an interim flag value missing, or a value of "Probably not a link" and date missing were deleted.

11 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight" and weights greater than or equal to 6, and sex and ethnicity agree, the following criteria were applied (in order):

- If country missing or agrees and, either, but not both, day or month disagree or missing, make a link.

- If pacific fix or asian fix agree, and either day and/or country disagree or missing, make a link.

12 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Look at - High Weight", the following criteria were applied (in order):

- If ethnicity agrees for Maori, Pacific and Asian (ignore nonMPA column in COMB variable) and year agrees, or day month and year tolerance agree, then make a link.

13 For records linked in an MB pass, with a weight greater than or equal to 4.99 and with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", the following criteria were applied (in order):

- If year is missing, but year tolerance is less than 5, and sex, day, month, country, Maori, Pacific, and Asian agree, then make a link.
- If date of birth all missing, but year tolerance less than 2 and sex, country and ethnicity agree, then make a link.
- If sex and year are agree, but ethnicity missing and country missing or agrees, then make a link.

14 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and year and year tolerance disagree (and country agrees or missing) the following criteria were applied (in order):

- If records linked in an MB pass and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and year tolerance is less than 2, then make a link.

15 For records with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Probably not a link" or "Look at - High Weight", and year missing and ethnicity (other than NonMPA part of COMB variable) agree or missing, the following criteria were applied (in order):

- If records linked in an MB pass and weight greater than 4.99 and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and weight greater than 6.49 and year tolerance is less than 3, then make a link.

16 For records with interim flag value "Look at - High Weight", and weight greater than or equal to 9, and year tolerance less than 2, then make a link.

17 For records with interim flag value missing, or values of "Strong possible link", or " Investigate possible link" or "Probably not a link" or "Look at - High Weight", the following criteria were applied (in order):

- If weight less than or equal to 4.99 then delete.
- If records linked in an MB pass and year tolerance is greater than 5, then delete.
- If records linked in an AU pass and (weight greater than 6.49 or year tolerance is greater than 1), then delete.

18 For the remaining records, assign the value of the final flag:

- For records with interim flag value of "Investigate possible link", then make a link.
- For records with interim flag value missing, or values of "Probably not a link" or "Look at - High Weight", then delete.

Table 1996.8.1 Summary of results of clerical review

Type of pass	Number linked in pass	Cumulative number linked (including final job)	% of cancer files
MB	3847	68484	71.01%
AU	8881	77365	80.22%

1996.9 Summary of links

Table 1996.9.1 Summary of links for 1996

Pass	Pass details	Number linked in pass	Cumulative number linked	Cumulative % of cancer files	PPV
1	MB1	47502	47502	49.25%	99.6
2	MB2	2102	49604	51.43%	96.4
3	AU1, sex	9058	58662	60.83%	83.7
4	AU2, sex	5975	64637	67.02%	83.5
CR MB	MB1 – MB3, MB2 (with AU2), MB3 (with AU3)	3847	68484	71.01%	
CR AU	AU1 – AU4, AU2 (after MB2), AU3 (after MB3)	8881	77365	80.22%	
Dups	Delete duplicates with identical weights	(474)	76891	79.73%	

CR MB – Clerical review records added in meshblock passes.
CR AU – Clerical review records added in area unit passes.
PPVs are not available for the clerical review passes.

1996.10 List of QualityStage reports for each job and pass

1996.10.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 24 July 2007

MatchOn123MB_1996_1.out
MatchOn123MB_1996_2.out
MatchOn123MB_1996_3.out

1996.10.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 25 July 2007

MatchOnMB12_3to6AU_1996_1.out
MatchOnMB12_3to6AU_1996_2.out
MatchOnMB12_3to6AU_1996_3.out
MatchOnMB12_3to6AU_1996_4.out
MatchOnMB12_3to6AU_1996_5.out
MatchOnMB12_3to6AU_1996_6.out

1996.10.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 27 July 2007

MatchOnMB12wco11pt99_34AUwco14pt6_1996_1.out
MatchOnMB12wco11pt99_34AUwco14pt6_1996_2.out
MatchOnMB12wco11pt99_34AUwco14pt6_1996_3.out
MatchOnMB12wco11pt99_34AUwco14pt6_1996_4.out

1996.10.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 1 August 2007

MatchOnRESp4MB123NCO_1996_1.out
MatchOnRESp4MB123NCO_1996_2.out
MatchOnRESp4MB123NCO_1996_3.out

1996.10.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 1 August 2007

MatchOnRESp4AU1234NCO_1996_1.out
MatchOnRESp4AU1234NCO_1996_2.out
MatchOnRESp4AU1234NCO_1996_3.out
MatchOnRESp4AU1234NCO_1996_4.out

1996.10.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 13 August 2007

MatchOnRESp4MB2AU2NOCO_1996_1.out
MatchOnRESp4MB2AU2NOCO_1996_2.out

1996.10.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 14 September 2007

MatchOnRESp4MB3AU3NOCO_1996_1.out
MatchOnRESp4MB3AU3NOCO_1996_2.out

1996.11 List of PPV calculations for each job and pass

1996.11.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 24 July 2007

Passes123MB NCO_1996_lastppvPass1.out
Passes123MB NCO_1996_lastppvPass2.out
Passes123MB NCO_1996_lastppvPass3.out

1996.11.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 25 July 2007

Passes12MBwco11pt99_3to6AUnco_1996_lastppvPass1.out
Passes12MBwco11pt99_3to6AUnco_1996_lastppvPass2.out
Passes12MBwco11pt99_3to6AUnco_1996_lastppvPass3.out
Passes12MBwco11pt99_3to6AUnco_1996_lastppvPass4.out
Passes12MBwco11pt99_3to6AUnco_1996_lastppvPass5.out
Passes12MBwco11pt99_3to6AUnco_1996_lastppvPass6.out

1996.11.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 27 July 2007

Passes12MBwco11pt99_34AUwco14pt60_1996_lastppvPass3.out
Passes12MBwco11pt99_34AUwco14pt60_1996_lastppvPass4.out

1991 Cohort

1991.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 26 July 2007

Table 1991.1.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1991.1.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3373927	MATCH	45711
	B Cancer	77159	DUPA	16847
			DUPB	164
			RESA	3311369
			RESB	31284
2 Block on MB2	RESA	3311369	MATCH	3081
	RESB	31284	DUPA	947
			DUPB	3
			RESA	3307341
			RESB	28200
3 Block on MB3	RESA	3307341	MATCH	277
	RESB	28200	DUPA	75
			DUPB	0
			RESA	3306989
			RESB	27923
Summary	A CTCensus	3373927	LINK1	49069
	B Cancer	77159	DUPA	17869
			DUPB	167
			RESA	3306989
			RESB	27923

1991.1.1 Decisions

1 Pass three, blocking on MB3W from the cancer file should not be used because there were only a small number of links (277) and the weights were too low. Good links can be picked up in clerical review.

2 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 12.99.

3 Four area unit passes, with no cut-off value should be run.

1991.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 27 July 2007

Table 1991.2.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	12.99
2	MB	MB2W	12.99
3	AU and SEX	AU1W and SEXW	0
4	AU and SEX	AU2W and SEXW	0
5	AU and SEX	AU3W and SEXW	0
6	AU and SEX	AU4W and SEXW	0

Table 1991.2.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3373927	MATCH	35853
	B Cancer	77159	DUPA	139
			DUPB	24
			RESA	3337935
			RESB	41282
2 Block on MB2	RESA	3337935	MATCH	2023
	RESB	41282	DUPA	8
			DUPB	1
			RESA	3335904
			RESB	39258
3 Block on AU1	RESA	3335904	MATCH	36679
	RESB	39258	DUPA	182605
			DUPB	75
			RESA	3116620
			RESB	2504
4 Block on AU2	RESA	3116620	MATCH	2021
	RESB	2504	DUPA	7392
			DUPB	3
			RESA	3107207
			RESB	480
5 Block on AU3	RESA	3107207	MATCH	247
	RESB	480	DUPA	1017
			DUPB	0
			RESA	3105943
			RESB	233
6 Block on AU4	RESA	3105943	MATCH	30
	RESB	233	DUPA	105
			DUPB	0
			RESA	3105808
			RESB	203

Summary	A CTCensus	3373927	LINK1	76853
	B Cancer	77159	DUPA	191266
			DUPB	103
			RESA	3105808
			RESB	203

1991.2.1 Decisions

1 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 12.99.

2 Passes 3 and 4 (blocking on AU1W and AU2W) should be rerun, with a cut-off value of 14.60.

3 Passes 5 and 6 (blocking on AU3W and AU4W on cancer file) should not be used because it produces only a small number of links (247 and 30) and the weights were too low. Some links can be picked up in clerical review.

1991.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 1 August 2007

Table 1991.3.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	12.99
2	MB	MB2W	12.99
3	AU and SEX	AU1W and SEXW	14.60
4	AU and SEX	AU2W and SEXW	14.60

Table 1991.3.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3373927	MATCH	35853
	B Cancer	77159	DUPA	139
			DUPB	24
			RESA	3337935
			RESB	41282
2 Block on MB2	RESA	3337935	MATCH	2023
	RESB	41282	DUPA	8
			DUPB	1
			RESA	3335904
			RESB	39258
3 Block on AU1	RESA	3335904	MATCH	8026
	RESB	39258	DUPA	279
			DUPB	17
			RESA	3327599
			RESB	31215
4 Block on AU2	RESA	3327599	MATCH	5709
	RESB	31215	DUPA	198
			DUPB	5
			RESA	3321692
			RESB	25501
Summary	A CTCensus	3373927	LINK1	51611
	B Cancer	77159	DUPA	624
			DUPB	47
			RESA	3321692
			RESB	25501

1991.3.1 Decisions

1 Links from this job will be included in the final linked file. (51611 linked files, or 66.89% of cancer files.)

2 Use residuals from this job (that is 3321692 census residuals and 25501 cancer residuals) for the following four jobs to produce records for clerical review. These jobs were considered the most likely to produce some good links that could be included in the final linked file.

3 Job 4, three passes blocking on MB1 to MB3 with no cut-off.

4 Job 5, four passes blocking on AU1 to AU4 with no cut-off.

5 Job 6, two passes blocking on MB2 and AU2 with no cut-off.

6 Job 7, two passes blocking on MB3 and AU3 with no cut-off.

1991.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 6 August 2007

Table 1991.4.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1991.4.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	RESA1 (Census)	3321692	MATCH	6816
	RESB1 (Cancer)	25501	DUPA	2465
			DUPB	13
			RESA	3312411
			RESB	18672
2 Block on MB2	RESA	3312411	MATCH	1178
	RESB	18672	DUPA	331
			DUPB	1
			RESA	3310902
			RESB	17493
3 Block on MB3	RESA	3310902	MATCH	163
	RESB	17493	DUPA	38
			DUPB	0
			RESA	3310701
			RESB	17330
Summary	RESA1 (Census)	3321692	LINK1	8157
	RESB1 (Cancer)	25501	DUPA	2834
			DUPB	14
			RESA	3310701
			RESB	17330

1991.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 6 August 2007

Table 1991.5.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	AU and SEX	AU1W and SEXW	0
2	AU and SEX	AU2W and SEXW	0
3	AU and SEX	AU3W and SEXW	0
4	AU and SEX	AU4W and SEXW	0

Table 1991.5.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on AU1	RESA1 (Census)	3321692	MATCH	23437
	RESB1 (Cancer)	25501	DUPA	117691
			DUPB	42
			RESA	3180564
			RESB	2022
2 Block on AU2	RESA	3180564	MATCH	1556
	RESB	2022	DUPA	5766
			DUPB	2
			RESA	3173242
			RESB	464
3 Block on AU3	RESA	3173242	MATCH	241
	RESB	464	DUPA	1017
			DUPB	0
			RESA	3171984
			RESB	223
4 Block on AU4	RESA	3171984	MATCH	29
	RESB	223	DUPA	101
			DUPB	0
			RESA	3171854
			RESB	194
Summary	RESA1 (Census)	3321692	LINK1	25263
	RESB1 (Cancer)	25501	DUPA	124575
			DUPB	44
			RESA	3171854
			RESB	194

1991.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 15 August 2007

Table 1991.6.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB2W	0
2	AU and SEX	AU2W and SEXW	0

Table 1991.6.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB2	RESA1 (Census)	3321692	MATCH	1704
	RESB1 (Cancer)	25501	DUPA	550
			DUPB	2
			RESA	3319438
			RESB	23795
2 Block on AU2	RESA	3319438	MATCH	17651
	RESB	23795	DUPA	90013
			DUPB	22
			RESA	3211774
			RESB	6122
Summary	RESA1 (Census)	3321692	LINK1	19355
	RESB1 (Cancer)	25501	DUPA	90563
			DUPB	24
			RESA	3211774
			RESB	6122

1991.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 10 September 2007

Table 1991.7.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB3W	0
2	AU and SEX	AU3W and SEXW	0

Table 1991.7.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB3	RESA1 (Census)	3321692	MATCH	301
	RESB1 (Cancer)	25501	DUPA	87
			DUPB	0
			RESA	3321304
			RESB	25200
2 Block on AU3	RESA	3321304	MATCH	11884
	RESB	25200	DUPA	67576
			DUPB	6
			RESA	3241844
			RESB	13310
Summary	RESA1 (Census)	3321692	LINK1	12185
	RESB1 (Cancer)	25501	DUPA	67663
			DUPB	6
			RESA	3241844
			RESB	13310

1991.8 Clerical review process

1991.8.1 Criteria used to include or exclude clerically reviewed records

Applying the criteria to decide if a clerically reviewed record should be included or excluded was an iterative process and in some cases resulted in attaching an interim flag to some records. The values of the interim flag were used to decide if a record should be finally included or excluded when later criteria were applied. A final flag was then attached to each record to indicate if it should be included or excluded from the final file.

Values of interim flag:

- Not a link
- Investigate possible link
- Probably not a link
- Look at - High Weight
- Make a link - Investigated & acceptable

In these criteria, records given an interim or final flag of "Not a link" are described as deleted. Records given an interim or final flag value of "Make a link - Investigated & acceptable" are described with the phrase "make a link".

Values of final flag:

- Not a link
- Make a link - Investigated & acceptable

Many of these criteria use parts of the COMB variable (eg day or month). A full description of the COMB variable is given in section 5.3.

In these criteria the phrase "year tolerance" is used in two different ways. First, as part of the COMB variable "year tolerance agrees" means that there is no difference in the year of birth between the two linked records or the difference is one year. Similarly, "Year tolerance disagrees" means that the difference is more than one year. Secondly, the YEARTOL variable measures the actual difference in years and the use of this variable is indicated in phrases like "year tolerance greater then 5". See sections 5.3 and 5.4 for fuller descriptions of the COMB and YEARTOL variables.

Criteria used to include or exclude clerically reviewed records:

1 Records with a weight below 5 were deleted.

2 Records linked in an AU pass with a weight below 6 were deleted.

3 Duplicate records with the same weight were deleted.

4 Records with a weight of 9 or more were flagged as "Look at - High Weight".

5 For records linked in an MB pass, with weights between 5 and 5.49, the following criteria were applied (in order):

- If year and year tolerance part of COMB variable indicate missing values then delete.
- If value of year tolerance variable greater than 5 then delete.
- If ethnicity values (Empan columns of COMB variable) disagree then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagree or year disagree or country of birth disagree then delete.
- else flag as "Probably not a link".

6 For records linked in an AU pass, with weights between 6 and 6.49, following criteria were applied (in order):

- If year and year tolerance missing then delete.
- If year tolerance greater than 1 then delete.
- If ethnicity (Empan) disagrees then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagrees or year disagrees, or country of birth disagrees then delete.
- else flag as "Probably not a link".

7 For records linked in an MB pass, the following criteria were applied (in order):

- For weights greater than or equal to 9 where day and month agree, or year or year tolerance agree, make a link.
- For weights with interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" and month, year and ethnicity agree, make a link.
- For weights greater than or equal to 9 and date of birth agrees (day, month and year) or day, month and year tolerance, or month and year agree, or day and year agree, or day and month and ethnicity agree, then make a link.

8 For records linked in an MB pass, where ethnicity agrees, and interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" the following criteria were applied (in order):

- If month and year agree, make a link.
- If day, month and year tolerance agree, make a link.
- If day and year agree, make a link.

9 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and ethnicity agrees or missing, make a link.

10 Records with an interim flag value missing, or a value of "Probably not a link" and date missing were deleted.

11 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight" and weights greater than or equal to 6, and sex and ethnicity agree, the following criteria were applied (in order):

- If country missing or agrees and, either, but not both, day or month disagree or missing, make a link.

- If pacific fix or asian fix agree, and either day and/or country disagree or missing, make a link.

12 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Look at - High Weight", the following criteria were applied (in order):

- If ethnicity agrees for Maori, Pacific and Asian (ignore nonMPA column in COMB variable) and year agrees, or day month and year tolerance agree, then make a link.

13 For records linked in an MB pass, with a weight greater than or equal to 5.49 and with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", the following criteria were applied (in order):

- If year is missing, but year tolerance is less than 5, and sex, day, month, country, Maori, Pacific, and Asian agree, then make a link.
- If date of birth all missing, but year tolerance less than 2 and sex, country and ethnicity agree, then make a link.
- If sex and year are agree, but ethnicity missing and country missing or agrees, then make a link.

14 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and year and year tolerance disagree (and country agrees or missing) the following criteria were applied (in order):

- If records linked in an MB pass and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and year tolerance is less than 2, then make a link.

15 For records with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Probably not a link" or "Look at - High Weight", and year missing and ethnicity (other than NonMPA part of COMB variable) agree or missing, the following criteria were applied (in order):

- If records linked in an MB pass and weight greater than 5.49 and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and weight greater than 6.49 and year tolerance is less than 3, then make a link.

16 For records with interim flag value "Look at - High Weight", and weight greater than or equal to 9, and year tolerance less than 2, then make a link.

17 For records with interim flag value missing, or values of "Strong possible link", or " Investigate possible link" or "Probably not a link" or "Look at - High Weight", the following criteria were applied (in order):

- If weight less than or equal to 5.49 then delete.
- If records linked in an MB pass and year tolerance is greater than 5, then delete.
- If records linked in an AU pass and (weight greater than 6.49 or year tolerance is greater than 1), then delete.

18 For the remaining records, assign the value of the final flag:

- For records with interim flag value of "Investigate possible link", then make a link.
- For records with interim flag value missing, or values of "Probably not a link" or "Look at - High Weight", then delete.

Table 1991.8.1 Summary of results of clerical review

Type of pass	Number linked in pass	Cumulative number linked (including final job)	% of cancer files
MB	2248	53859	69.80%
AU	7628	61487	79.69%

1991.9 Summary of links

Table 1991.9.1 Summary of links for 1991

Pass	Pass details	Number linked in pass	Cumulative number linked	Cumulative % of cancer files	PPV
1	MB1	35853	35853	46.47%	99.7
2	MB2	2023	37876	49.09%	98.3
3	AU1, sex	8026	45902	59.49%	85.3
4	AU2, sex	5709	51611	66.89%	78.9
CR MB	MB1 – MB3, MB2 (with AU2), MB3 (with AU3)	2248	53859	69.80%	
CR AU	AU1 – AU4, AU2 (after MB2), AU3 (after MB3)	7628	61487	79.69%	
Dups	Delete duplicates with identical weights	(382)	61105	79.19%	

CR MB – Clerical review records added in meshblock passes.
CR AU – Clerical review records added in area unit passes.
PPVs are not available for the clerical review passes.

1991.10 List of QualityStage reports for each job and pass

1991.10.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 26 July 2007

MatchOnMB123wnco_1991_1.out
MatchOnMB123wnco_1991_2.out
MatchOnMB123wnco_1991_3.out

1991.10.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 27 July 2007

MatchOnMB12wco12pt99_3to6AUwnco_1991_1.out
MatchOnMB12wco12pt99_3to6AUwnco_1991_2.out
MatchOnMB12wco12pt99_3to6AUwnco_1991_3.out
MatchOnMB12wco12pt99_3to6AUwnco_1991_4.out
MatchOnMB12wco12pt99_3to6AUwnco_1991_5.out
MatchOnMB12wco12pt99_3to6AUwnco_1991_6.out

1991.10.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 1 August 2007

MatchOnMB12wco12pt99_34AUwco14pt6_1991_1.out
MatchOnMB12wco12pt99_34AUwco14pt6_1991_2.out
MatchOnMB12wco12pt99_34AUwco14pt6_1991_3.out
MatchOnMB12wco12pt99_34AUwco14pt6_1991_4.out

1991.10.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 6 August 2007

MatchOnRESp4MB123NCO_1991_1.out
MatchOnRESp4MB123NCO_1991_2.out
MatchOnRESp4MB123NCO_1991_3.out

1991.10.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 6 August 2007

MatchOnRESp4AU1234NCO_1991_1.out
MatchOnRESp4AU1234NCO_1991_2.out
MatchOnRESp4AU1234NCO_1991_3.out
MatchOnRESp4AU1234NCO_1991_4.out

1991.10.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 15 August 2007

MatchOnRESp4MB2AU2NOCO_1991_1.out
MatchOnRESp4MB2AU2NOCO_1991_2.out

1991.10.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off - 10 September 2007

MatchOnRESp4MB3AU3NCO_1991_1.out
MatchOnRESp4MB3AU3NCO_1991_2.out

1991.11 List of PPV calculations for each job and pass

1991.11.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 26 July 2007

Passes123MB NCO_1991_lastppvPass1.out
Passes123MB NCO_1991_lastppvPass2.out
Passes123MB NCO_1991_lastppvPass3.out

1991.11.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 27 July 2007

Passes12MBwco12pt99_3to6AUnco_1991_lastppvPass1.out
Passes12MBwco12pt99_3to6AUnco_1991_lastppvPass2.out
Passes12MBwco12pt99_3to6AUnco_1991_lastppvPass3.out
Passes12MBwco12pt99_3to6AUnco_1991_lastppvPass4.out
Passes12MBwco12pt99_3to6AUnco_1991_lastppvPass5.out
Passes12MBwco12pt99_3to6AUnco_1991_lastppvPass6.out

1991.11.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.60 – 1 August 2007

Passes12MBwco12pt99_34AUwco14pt60_1991_lastppvPass3.out
Passes12MBwco12pt99_34AUwco14pt60_1991_lastppvPass4.out

1986 Cohort

1986.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 7 August 2007

Table 1986.1.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1986.1.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3263284	MATCH	42398
	B Cancer	63626	DUPA	21677
			DUPB	789
			RESA	3199209
			RESB	20439
2 Block on MB2	RESA	3199209	MATCH	2998
	RESB	20439	DUPA	1272
			DUPB	21
			RESA	3194939
			RESB	17420
3 Block on MB3	RESA	3194939	MATCH	393
	RESB	17420	DUPA	138
			DUPB	1
			RESA	3194408
			RESB	17026
Summary	A CTCensus	3263284	LINK1	45789
	B Cancer	63626	DUPA	23087
			DUPB	811
			RESA	3194408
			RESB	17026

1986.1.1 Decisions

1 Pass three, blocking on MB3W from the cancer file should not be used because there were only a small number of links (393) and the weights were too low. Good links can be picked up in clerical review.

2 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 12.99.

3 Four area unit passes, with no cut-off value should be run.

1986.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 7 August 2007

Table 1986.2.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	12.99
2	MB	MB2W	12.99
3	AU and SEX	AU1W and SEXW	0
4	AU and SEX	AU2W and SEXW	0
5	AU and SEX	AU3W and SEXW	0
6	AU and SEX	AU4W and SEXW	0

Table 1986.2.2 Summary of match

Pass number and description	In		Out	
1 Block on MB1	A CTCensus	3263284	MATCH	30998
	B Cancer	63626	DUPA	129
			DUPB	14
			RESA	3232157
			RESB	32614
2 Block on MB2	RESA	3232157	MATCH	2339
	RESB	32614	DUPA	10
			DUPB	0
			RESA	3229808
			RESB	30275
3 Block on AU1	RESA	3229808	MATCH	28978
	RESB	30275	DUPA	154632
			DUPB	499
			RESA	3046198
			RESB	798
4 Block on AU2	RESA	3046198	MATCH	621
	RESB	798	DUPA	2222
			DUPB	8
			RESA	3043355
			RESB	169
5 Block on AU3	RESA	3043355	MATCH	90
	RESB	169	DUPA	380
			DUPB	0
			RESA	3042885
			RESB	79
6 Block on AU4	RESA	3042885	MATCH	5
	RESB	79	DUPA	22
			DUPB	0
			RESA	3042858

			RESB	74
Summary	A CTCensus	3263284	LINK1	63031
	B Cancer	63626	DUPA	157395
			DUPB	521
			RESA	3042858
			RESB	74

1986.2.1 Decisions

1 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 12.99.

2 Passes 3 and 4 (blocking on AU1W and AU2W) should be rerun, with a cut-off value of 14.75.

3 Passes 5 and 6 (blocking on AU3W and AU4W) should not be used because they produced only a small number of links (90 and five respectively) and the weights were too low. Good links can be picked up in clerical review.

1986.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.75 – 8 August 2007

Table 1986.3.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	12.99
2	MB	MB2W	12.99
3	AU and SEX	AU1W and SEXW	14.75
4	AU and SEX	AU2W and SEXW	14.75

Table 1986.3.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3263284	MATCH	30998
	B Cancer	63626	DUPA	129
			DUPB	14
			RESA	3232157
			RESB	32614
2 Block on MB2	RESA	3232157	MATCH	2339
	RESB	32614	DUPA	10
			DUPB	0
			RESA	3229808
			RESB	30275
3 Block on AU1	RESA	3229808	MATCH	4870
	RESB	30275	DUPA	166
			DUPB	9
			RESA	3224772
			RESB	25396
4 Block on AU2	RESA	3224772	MATCH	2956
	RESB	25396	DUPA	93
			DUPB	5
			RESA	3221723
			RESB	22435
Summary	A CTCensus	3263284	LINK1	41163
	B Cancer	63626	DUPA	398
			DUPB	28
			RESA	3221723
			RESB	22435

1986.3.1 Decisions

1 Links from this job will be included in the final linked file. (41163 linked files, or 64.70% of cancer files)

2 Use residuals from this job (that is 3221723 census residuals and 22435 cancer residuals) for four jobs to produce records for clerical review. These jobs were considered the most likely to produce some good links that could be included in the final linked file.

3 Job 4, three passes blocking on MB1 to MB3 with no cut-off.

4 Job 5, four passes blocking on AU1 to AU4 with no cut-off.

5 Job 6, two passes blocking on MB2 and AU2 with no cut-off.

6 Job 7, two passes blocking on MB3 and AU3 with no cut-off.

1986.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 8 August 2007

Table 1986.4.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1986.4.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	RESA1 (Census)	3221723	MATCH	8663
	RESB1 (Cancer)	22435	DUPA	5409
			DUPB	349
			RESA	3207651
			RESB	13423
2 Block on MB2	RESA	3207651	MATCH	1466
	RESB	13423	DUPA	694
			DUPB	17
			RESA	3205491
			RESB	11940
3 Block on MB3	RESA	3205491	MATCH	298
	RESB	11940	DUPA	123
			DUPB	0
			RESA	3205070
			RESB	11642
Summary	RESA1 (Census)	3221723	LINK1	10427
	RESB1 (Cancer)	22435	DUPA	6226
			DUPB	366
			RESA	3205070
			RESB	11642

1986.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 9 August 2007

Table 1986.5.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	AU and SEX	AU1W and SEXW	0
2	AU and SEX	AU2W and SEXW	0
3	AU and SEX	AU3W and SEXW	0
4	AU and SEX	AU4W and SEXW	0

Table 1986.5.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on AU1	RESA1 (Census)	3221723	MATCH	21354
	RESB1 (Cancer)	22435	DUPA	116879
			DUPB	390
			RESA	3083490
			RESB	691
2 Block on AU2	RESA	3083490	MATCH	517
	RESB	691	DUPA	1809
			DUPB	8
			RESA	3081164
			RESB	166
3 Block on AU3	RESA	3081164	MATCH	88
	RESB	166	DUPA	388
			DUPB	0
			RESA	3080688
			RESB	78
4 Block on AU4	RESA	3080688	MATCH	5
	RESB	78	DUPA	22
			DUPB	0
			RESA	3080661
			RESB	73
Summary	RESA1 (Census)	3221723	LINK1	21964
	RESB1 (Cancer)	22435	DUPA	119098
			DUPB	398
			RESA	3080661
			RESB	73

1986.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 15 August 2007

Table 1986.6.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB2W	0
2	AU and SEX	AU2W and SEXW	0

Table 1986.6.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB2	RESA1 (Census)	3221723	MATCH	3056
	RESB1 (Cancer)	22435	DUPA	1845
			DUPB	96
			RESA	3216822
			RESB	19283
2 Block on AU2	RESA	3216822	MATCH	13531
	RESB	19283	DUPA	80393
			DUPB	224
			RESA	3122898
			RESB	5528
Summary	RESA1 (Census)	3221723	LINK1	16587
	RESB1 (Cancer)	22435	DUPA	82238
			DUPB	320
			RESA	3122898
			RESB	5528

1986.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 10 September 2007

Table 1986.7.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB3W	0
2	AU and SEX	AU3W and SEXW	0

Table 1986.7.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB3	RESA1 (Census)	3221723	MATCH	871
	RESB1 (Cancer)	22435	DUPA	597
			DUPB	7
			RESA	3220255
			RESB	21557
2 Block on AU3	RESA	3220255	MATCH	9104
	RESB	21557	DUPA	63511
			DUPB	111
			RESA	3147640
			RESB	12342
Summary	RESA1 (Census)	3221723	LINK1	9975
	RESB1 (Cancer)	22435	DUPA	64108
			DUPB	118
			RESA	3147640
			RESB	12342

1986.8 Clerical review process

1986.8.1 Criteria used to include or exclude clerically reviewed records

Applying the criteria to decide if a clerically reviewed record should be included or excluded was an iterative process and in some cases resulted in attaching an interim flag to some records. The values of the interim flag were used to decide if a record should be finally included or excluded when later criteria were applied. A final flag was then attached to each record to indicate if it should be included or excluded from the final file.

Values of interim flag:

- Not a link
- Investigate possible link
- Probably not a link
- Look at - High Weight
- Make a link - Investigated & acceptable

In these criteria, records given an interim or final flag of "Not a link" are described as deleted. Records given an interim or final flag value of "Make a link - Investigated & acceptable" are described with the phrase "make a link".

Values of final flag:

- Not a link
- Make a link - Investigated & acceptable

Many of these criteria use parts of the COMB variable (eg day or month). A full description of the COMB variable is given in section 5.3.

In these criteria the phrase "year tolerance" is used in two different ways. First, as part of the COMB variable "year tolerance agrees" means that there is no difference in the year of birth between the two linked records or the difference is one year. Similarly, "Year tolerance disagrees" means that the difference is more than one year. Secondly, the YEARTOL variable measures the actual difference in years and the use of this variable is indicated in phrases like "year tolerance greater then 5". See sections 5.3 and 5.4 for fuller descriptions of the COMB and YEARTOL variables.

Criteria used to include or exclude clerically reviewed records:

1 Records with a weight below 4 were deleted.

2 Records linked in an AU pass with a weight below 4.3 were deleted.

3 Duplicate records with the same weight were deleted.

4 Records with a weight of 9 or more were flagged as "Look at - High Weight".

5 For records linked in an MB pass, with weights between 4 and 4.99, the following criteria were applied (in order):

- If year and year tolerance part of COMB variable indicate missing values then delete.
- If value of year tolerance variable greater than 5 then delete.
- If ethnicity values (Empan columns of COMB variable) disagree then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagree or year disagree or country of birth disagree then delete.
- else flag as "Probably not a link".

6 For records linked in an AU pass, with weights between 4.3 and 6.49, following criteria were applied (in order):

- If year and year tolerance missing then delete.
- If year tolerance greater than 1 then delete.
- If ethnicity (Empan) disagrees then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagrees or year disagrees, or country of birth disagrees then delete.
- else flag as "Probably not a link".

7 For records linked in an MB pass, the following criteria were applied (in order):

- For weights greater than or equal to 9 where day and month agree, or year or year tolerance agree, make a link.
- For weights with interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" and month, year and ethnicity agree, make a link.
- For weights greater than or equal to 9 and date of birth agrees (day, month and year) or day, month and year tolerance, or month and year agree, or day and year agree, or day and month and ethnicity agree, then make a link.

8 For records linked in an MB pass, where ethnicity agrees, and interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" the following criteria were applied (in order):

- If month and year agree, make a link.
- If day, month and year tolerance agree, make a link.
- If day and year agree, make a link.

9 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and ethnicity agrees or missing, make a link.

10 Records with an interim flag value missing, or a value of "Probably not a link" and date missing were deleted.

11 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight" and weights greater than or equal to 6, and sex and ethnicity agree, the following criteria were applied (in order):

- If country missing or agrees and, either, but not both, day or month disagree or missing, make a link.

- If pacific fix or asian fix agree, and either day and/or country disagree or missing, make a link.

12 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Look at - High Weight", the following criteria were applied (in order):

- If ethnicity agrees for Maori, Pacific and Asian (ignore nonMPA column in COMB variable) and year agrees, or day month and year tolerance agree, then make a link.

13 For records linked in an MB pass, with a weight greater than or equal to 4.99 and with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", the following criteria were applied (in order):

- If year is missing, but year tolerance is less than 5, and sex, day, month, country, Maori, Pacific, and Asian agree, then make a link.
- If date of birth all missing, but year tolerance less than 2 and sex, country and ethnicity agree, then make a link.
- If sex and year are agree, but ethnicity missing and country missing or agrees, then make a link.

14 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and year and year tolerance disagree (and country agrees or missing) the following criteria were applied (in order):

- If records linked in an MB pass and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and year tolerance is less than 2, then make a link.

15 For records with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Probably not a link" or "Look at - High Weight", and year missing and ethnicity (other than NonMPA part of COMB variable) agree or missing, the following criteria were applied (in order):

- If records linked in an MB pass and weight greater than 4.99 and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and weight greater than 6.49 and year tolerance is less than 3, then make a link.

16 For records with interim flag value "Look at - High Weight", and weight greater than or equal to 9, and year tolerance less than 2, then make a link.

17 For records with interim flag value missing, or values of "Strong possible link", or " Investigate possible link" or "Probably not a link" or "Look at - High Weight", the following criteria were applied (in order):

- If weight less than or equal to 4.99 then delete.
- If records linked in an MB pass and year tolerance is greater than 5, then delete.
- If records linked in an AU pass and (weight greater than 6.49 or year tolerance is greater than 1), then delete.

18 For the remaining records, assign the value of the final flag:

- For records with interim flag value of "Investigate possible link", then make a link.
- For records with interim flag value missing, or values of "Probably not a link" or "Look at - High Weight", then delete.

Table 1986.8.1 Summary of results of clerical review

Type of pass	Number linked in pass	Cumulative number linked (including final job)	% of cancer files
MB	2664	43827	68.88%
AU	5438	49265	77.43%

1996.9 Summary of links

Table 1986.9.1 Summary of links for 1986

Pass	Pass details	Number linked in pass	Cumulative number linked	Cumulative % of cancer files	PPV
1	MB1	30998	30998	48.72%	99.6
2	MB2	2339	33337	52.40%	98.3
3	AU1, sex	4870	38207	60.05%	80.2
4	AU2, sex	2956	41163	64.70%	79.6
CR MB	MB1 – MB3, MB2 (with AU2), MB3 (with AU3)	2664	43827	68.88%	
CR AU	AU1 – AU4, AU2 (after MB2), AU3 (after MB3)	5438	49265	77.43%	
Dups	Delete duplicates with identical weights	(244)	49021	77.05%	

CR MB – Clerical review records added in meshblock passes.
CR AU – Clerical review records added in area unit passes.
PPVs are not available for the clerical review passes.

1986.10 List of QualityStage reports for each job and pass

1986.10.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off –7 August 2007

MatchOnMB123nco_1986_1.out
MatchOnMB123nco_1986_2.out
MatchOnMB123nco_1986_3.out

1986.10.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 7 August 2007

MatchOnMB12wco12pt99_3to6AUwnco_1986_1.out
MatchOnMB12wco12pt99_3to6AUwnco_1986_2.out
MatchOnMB12wco12pt99_3to6AUwnco_1986_3.out
MatchOnMB12wco12pt99_3to6AUwnco_1986_4.out
MatchOnMB12wco12pt99_3to6AUwnco_1986_5.out
MatchOnMB12wco12pt99_3to6AUwnco_1986_6.out

1986.10.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.75 – 8 August 2007

MatchOnMB12wco12pt99_3to4AUwco14pt75_1986_1.out
MatchOnMB12wco12pt99_3to4AUwco14pt75_1986_2.out
MatchOnMB12wco12pt99_3to4AUwco14pt75_1986_3.out
MatchOnMB12wco12pt99_3to4AUwco14pt75_1986_4.out

1986.10.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off– 8 August 2007

MatchOnRESp4MB123NCO_1986_1.out
MatchOnRESp4MB123NCO_1986_2.out
MatchOnRESp4MB123NCO_1986_3.out

1986.10.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 9 August 2007

MatchOnRESp4AU1234NCO_1986_1.out
MatchOnRESp4AU1234NCO_1986_2.out
MatchOnRESp4AU1234NCO_1986_3.out
MatchOnRESp4AU1234NCO_1986_4.out

1986.10.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 15 August 2007

MatchOnRESp4MB2AU2NCO_1986_1.out
MatchOnRESp4MB2AU2NCO_1986_2.out

1986.10.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 10 September 2007

MatchOnRESp4MB3AU3NCO_1986_1.out
MatchOnRESp4MB3AU3NCO_1986_2.out

1986.11 List of PPV calculations for each job and pass

1986.11.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off –7 August 2007

Passes123MB NCO_1986_lastppvPass1.out
Passes123MB NCO_1986_lastppvPass2.out
Passes123MB NCO_1986_lastppvPass3.out

1986.11.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 7 August 2007

Passes12MBwco12pt99_3to6AUnco_1986_lastppvPass1.out
Passes12MBwco12pt99_3to6AUnco_1986_lastppvPass2.out
Passes12MBwco12pt99_3to6AUnco_1986_lastppvPass3.out
Passes12MBwco12pt99_3to6AUnco_1986_lastppvPass4.out
Passes12MBwco12pt99_3to6AUnco_1986_lastppvPass5.out
Passes12MBwco12pt99_3to6AUnco_1986_lastppvPass6.out

1986.11.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 12.99, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 14.75 – 8 August 2007

Passes12MBwco12pt99_34AUwco14pt75_1986_lastppvPass3.out
Passes12MBwco12pt99_34AUwco14pt75_1986_lastppvPass4.out

1981 Cohort

1981.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 10 August 2007

Table 1981.1.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1981.1.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3143307	MATCH	30816
	B Cancer	52699	DUPA	10305
			DUPB	75
			RESA	3102186
			RESB	21808
2 Block on MB2	RESA	3102186	MATCH	1808
	RESB	21808	DUPA	523
			DUPB	5
			RESA	3099855
			RESB	19995
3 Block on MB3	RESA	3099855	MATCH	254
	RESB	19995	DUPA	67
			DUPB	0
			RESA	3099534
			RESB	19741
Summary	A CTCensus	3143307	LINK1	32878
	B Cancer	52699	DUPA	10895
			DUPB	80
			RESA	3099534
			RESB	19741

1981.1.1 Decisions

1 Pass three, blocking on MB3W from the cancer file should not be used because there were only a small number of links (254) and the weights were too low. Good links can be picked up in clerical review.

2 Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 11.67.

3 Four area unit passes, with no cut-off value should be run.

1981.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.67, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 13 August 2007

Table 1981.2.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	11.67
2	MB	MB2W	11.67
3	AU and SEX	AU1W and SEXW	0
4	AU and SEX	AU2W and SEXW	0
5	AU and SEX	AU3W and SEXW	0
6	AU and SEX	AU4W and SEXW	0

Table 1981.2.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	A CTCensus	3143307	MATCH	23780
	B Cancer	52699	DUPA	103
			DUPB	13
			RESA	3119424
			RESB	28906
2 Block on MB2	RESA	3119424	MATCH	1075
	RESB	28906	DUPA	5
			DUPB	0
			RESA	3118344
			RESB	27831
3 Block on AU1	RESA	3118344	MATCH	25782
	RESB	27831	DUPA	118945
			DUPB	73
			RESA	2973617
			RESB	1976
4 Block on AU2	RESA	2973617	MATCH	1253
	RESB	1976	DUPA	4383
			DUPB	4
			RESA	2967981
			RESB	719
5 Block on AU3	RESA	2967981	MATCH	142
	RESB	719	DUPA	554
			DUPB	0
			RESA	2967285
			RESB	577
6 Block on AU4	RESA	2967285	MATCH	19
	RESB	577	DUPA	39
			DUPB	0
			RESA	2967227

			RESB	558
Summary	A CTCensus	3143307	LINK1	52051
	B Cancer	52699	DUPA	124029
			DUPB	90
			RESA	2967227
			RESB	558

1981.2.1 Decisions

1. Passes 1 and 2 (blocking on MB1W and MB2W) should be rerun, with a cut-off value of 11.67.

2. Passes 3 and 4 (blocking on AU1W and AU2W) should be rerun, with a cut-off value of 13.65.

3. Passes 5 and 6 (blocking on AU3W and AU4W) should not be used because they produce only a small number of links (142 and 19) and the weights were too low. Good links can be picked up in clerical review.

1981.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.67, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 13.65 – 14 August 2007

Table 1981.3.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB1W	11.67
2	MB	MB2W	11.67
3	AU and SEX	AU1W and SEXW	13.65
4	AU and SEX	AU2W and SEXW	13.65

Table 1981.3.2 Summary of match

Pass number and description	**In**		**out**	
1 Block on MB1	A CTCensus	3143307	MATCH	23780
	B Cancer	52699	DUPA	103
			DUPB	13
			RESA	3119424
			RESB	28906
2 Block on MB2	RESA	3119424	MATCH	1075
	RESB	28906	DUPA	5
			DUPB	0
			RESA	3118344
			RESB	27831
3 Block on AU1	RESA	3118344	MATCH	5465
	RESB	27831	DUPA	185
			DUPB	9
			RESA	3112694
			RESB	22357
4 Block on AU2	RESA	3112694	MATCH	2668
	RESB	22357	DUPA	127
			DUPB	2
			RESA	3109899
			RESB	19687
Summary	A CTCensus	3143307	LINK1	32988
	B Cancer	52699	DUPA	420
			DUPB	24
			RESA	3109899
			RESB	19687

1981.3.1 Decisions

1 Links from this job will be included in the final linked file. (32988 linked files, or 62.60% of cancer files.)

2 Use residuals from this job (that is 3109899 census residuals and 19687 cancer residuals) for four jobs to produce records for clerical review. These jobs were considered the most likely to produce some good links that could be included in the final linked file.

3 Job 4, three passes blocking on MB1 to MB3 with no cut-off.

4 Job 5, four passes blocking on AU1 to AU4 with no cut-off.

5 Job 6, two passes blocking on MB2 and AU2 with no cut-off.

6 Job 7, two passes blocking on MB3 and AU3 with no cut-off.

1981.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 14 August 2007

Table 1981.4.1 Details of passes

Pass number	Block variable A	Block variable B	Cut-off
1	MB	MB1W	0
2	MB	MB2W	0
3	MB	MB3W	0

Table 1981.4.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB1	RESA1 (Census)	3109899	MATCH	5463
	RESB1 (Cancer)	19687	DUPA	1796
			DUPB	8
			RESA	3102640
			RESB	14216
2 Block on MB2	RESA	3102640	MATCH	743
	RESB	14216	DUPA	174
			DUPB	2
			RESA	3101723
			RESB	13471
3 Block on MB3	RESA	3101723	MATCH	159
	RESB	13471	DUPA	43
			DUPB	0
			RESA	3101521
			RESB	13312
Summary	RESA1 (Census)	3109899	LINK1	6365
	RESB1 (Cancer)	19687	DUPA	2013
			DUPB	10
			RESA	3101521
			RESB	13312

1981.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 14 August 2007

Table 1981.5.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	AU and SEX	AU1W and SEXW	0
2	AU and SEX	AU2W and SEXW	0
3	AU and SEX	AU3W and SEXW	0
4	AU and SEX	AU4W and SEXW	0

Table 1981.5.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on AU1	RESA1 (Census)	3109899	MATCH	17960
	RESB1 (Cancer)	19687	DUPA	80091
			DUPB	49
			RESA	3011848
			RESB	1678
2 Block on AU2	RESA	3011848	MATCH	959
	RESB	1678	DUPA	3135
			DUPB	4
			RESA	3007754
			RESB	715
3 Block on AU3	RESA	3007754	MATCH	141
	RESB	715	DUPA	562
			DUPB	0
			RESA	3007051
			RESB	574
4 Block on AU4	RESA	3007051	MATCH	19
	RESB	574	DUPA	39
			DUPB	0
			RESA	3006993
			RESB	555
Summary	RESA1 (Census)	3109899	LINK1	19079
	RESB1 (Cancer)	19687	DUPA	83827
			DUPB	53
			RESA	3006993
			RESB	555

1981.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 16 August 2007

Table 1981.6.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB2W	0
2	AU and SEX	AU2W and SEXW	0

Table 1981.6.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB2	RESA1 (Census)	3109899	MATCH	1102
	RESB1 (Cancer)	19687	DUPA	292
			DUPB	4
			RESA	3108505
			RESB	18581
2 Block on AU2	RESA	3108505	MATCH	9762
	RESB	18581	DUPA	44354
			DUPB	16
			RESA	3054389
			RESB	8803
Summary	RESA1 (Census)	3109899	LINK1	10864
	RESB1 (Cancer)	19687	DUPA	44646
			DUPB	20
			RESA	3054389
			RESB	8803

1981.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 5 September 2007

Table 1981.7.1 Details of passes

Pass number	Block variable(s) A	Block variable(s) B	Cut-off
1	MB	MB3W	0
2	AU and SEX	AU3W and SEXW	0

Table 1981.7.2 Summary of match

Pass number and description	**In**		**Out**	
1 Block on MB3	RESA1 (Census)	3109899	MATCH	280
	RESB1 (Cancer)	19687	DUPA	97
			DUPB	0
			RESA	3109522
			RESB	19407
2 Block on AU3	RESA	3109522	MATCH	4850
	RESB	19407	DUPA	23399
			DUPB	2
			RESA	3081273
			RESB	14555
Summary	RESA1 (Census)	3109899	LINK1	5130
	RESB1 (Cancer)	19687	DUPA	23496
			DUPB	2
			RESA	3081273
			RESB	14555

1981.8 Clerical review process

1981.8.1 Criteria used to include or exclude clerically reviewed records

Applying the criteria to decide if a clerically reviewed record should be included or excluded was an iterative process and in some cases resulted in attaching an interim flag to some records. The values of the interim flag were used to decide if a record should be finally included or excluded when later criteria were applied. A final flag was then attached to each record to indicate if it should be included or excluded from the final file.

Values of interim flag:

- Not a link
- Investigate possible link
- Probably not a link
- Look at - High Weight
- Make a link - Investigated & acceptable

In these criteria, records given an interim or final flag of "Not a link" are described as deleted. Records given an interim or final flag value of "Make a link - Investigated & acceptable" are described with the phrase "make a link".

Values of final flag:

- Not a link
- Make a link - Investigated & acceptable

Many of these criteria use parts of the COMB variable (eg day or month). A full description of the COMB variable is given in section 5.3.

In these criteria the phrase "year tolerance" is used in two different ways. First, as part of the COMB variable "year tolerance agrees" means that there is no difference in the year of birth between the two linked records or the difference is one year. Similarly, "Year tolerance disagrees" means that the difference is more than one year. Secondly, the YEARTOL variable measures the actual difference in years and the use of this variable is indicated in phrases like "year tolerance greater then 5". See sections 5.3 and 5.4 for fuller descriptions of the COMB and YEARTOL variables.

Criteria used to include or exclude clerically reviewed records:

1 Records with a weight below 4 were deleted.

2 Records linked in an AU pass with a weight below 4.3 were deleted.

3 Duplicate records with the same weight were deleted.

4 Records with a weight of 8 or more were flagged as "Look at - High Weight".

5 For records linked in an MB pass, with weights between 4 and 4.99, the following criteria were applied (in order):

- If year and year tolerance part of COMB variable indicate missing values then delete.
- If value of year tolerance variable greater than 5 then delete.
- If ethnicity values (Empan columns of COMB variable) disagree then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagree or year disagree or country of birth disagree then delete.
- else flag as "Probably not a link".

6 For records linked in an AU pass, with weights between 4.3 and 6.49, following criteria were applied (in order):

- If year and year tolerance missing then delete.
- If year tolerance greater than 1 then delete.
- If ethnicity (Empan) disagrees then delete.
- else if only year, only month, or only day disagree or missing but rest of date and ethnicity agree and country of birth either agrees or missing then flag as "investigate possible link". (Can't have date and country both missing.)
- else if day and month disagree, or day disagrees or year disagrees, or country of birth disagrees then delete.
- else flag as "Probably not a link".

7 For records linked in an MB pass, the following criteria were applied (in order):

- For weights greater than or equal to 8 where day and month agree, or year or year tolerance agree, make a link.
- For weights with interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" and month, year and ethnicity agree, make a link.
- For weights greater than or equal to 8 and date of birth agrees (day, month and year) or day, month and year tolerance, or month and year agree, or day and year agree, or day and month and ethnicity agree, then make a link.

8 For records linked in an MB pass, where ethnicity agrees, and interim flag value missing, or values of "Strong possible link", "Investigate possible link" or "Look at - High Weight" the following criteria were applied (in order):

- If month and year agree, make a link.
- If day, month and year tolerance agree, make a link.
- If day and year agree, make a link.

9 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and ethnicity agrees or missing, make a link.

10 Records with an interim flag value missing, or a value of "Probably not a link" and date missing were deleted.

11 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight" and weights greater than or equal to 6, and sex and ethnicity agree, the following criteria were applied (in order):

- If country missing or agrees and, either, but not both, day or month disagree or missing, make a link.

- If pacific fix or asian fix agree, and either day and/or country disagree or missing, make a link.

12 For records linked in an MB pass, with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Look at - High Weight", the following criteria were applied (in order):

- If ethnicity agrees for Maori, Pacific and Asian (ignore nonMPA column in COMB variable) and year agrees, or day month and year tolerance agree, then make a link.

13 For records linked in an MB pass, with a weight greater than or equal to 4.99 and with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", the following criteria were applied (in order):

- If year is missing, but year tolerance is less than 5, and sex, day, month, country, Maori, Pacific, and Asian agree, then make a link.
- If date of birth all missing, but year tolerance less than 2 and sex, country and ethnicity agree, then make a link.
- If sex and year are agree, but ethnicity missing and country missing or agrees, then make a link.

14 For records with interim flag value missing, or values of "Strong possible link", or "Look at - High Weight", and year and year tolerance disagree (and country agrees or missing) the following criteria were applied (in order):

- If records linked in an MB pass and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and year tolerance is less than 2, then make a link.

15 For records with interim flag value missing, or values of "Strong possible link", or "Investigate possible link" or "Probably not a link" or "Look at - High Weight", and year missing and ethnicity (other than NonMPA part of COMB variable) agree or missing, the following criteria were applied (in order):

- If records linked in an MB pass and weight greater than 4.99 and year tolerance is less than 6, then make a link.
- If records linked in an AU pass and weight greater than 6.49 and year tolerance is less than 3, then make a link.

16 For records with interim flag value "Look at - High Weight", and weight greater than or equal to 8, and year tolerance less than 2, then make a link.

17 For records with interim flag value missing, or values of "Strong possible link", or " Investigate possible link" or "Probably not a link" or "Look at - High Weight", the following criteria were applied (in order):

- If weight less than or equal to 4.99 then delete.
- If records linked in an MB pass and year tolerance is greater than 5, then delete.
- If records linked in an AU pass and (weight greater than 6.49 or year tolerance is greater than 1), then delete.

18 For the remaining records, assign the value of the final flag:

- For records with interim flag value of "Investigate possible link", then make a link.
- For records with interim flag value missing, or values of "Probably not a link" or "Look at - High Weight", then delete.

Table 1981.8.1 Summary of results of clerical review

Type of pass	Number linked in pass	Cumulative number linked (including final job)	% of cancer files
MB	2367	35355	67.09%
AU	3437	38792	73.61%

1981.9 Summary of links

Table 1981.9.1 Summary of links for 1981

Pass	Pass details	Number linked in pass	Cumulative number linked	Cumulative % of cancer files	PPV
1	MB1	23780	23780	45.12%	99.6
2	MB2	1075	24855	47.16%	97.4
3	AU1, sex	5465	30320	57.53%	84.9
4	AU2, sex	2668	32988	62.60%	75.8
CR MB	MB1 – MB3, MB2 (with AU2), MB3 (with AU3)	2367	35355	67.09%	
CR AU	AU1 – AU4, AU2 (after MB2), AU3 (after MB3)	3437	38792	73.61%	
Dups	Delete identical duplicates	(229)	38563	73.18%	

CR MB – Clerical review records added in meshblock passes.
CR AU – Clerical review records added in area unit passes.
PPVs are not available for the clerical review passes.

1981.10 List of QualityStage reports for each job and pass

1981.10.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 10 August 2007

MatchOnMB123wnco_1981_1.out
MatchOnMB123wnco_1981_2.out
MatchOnMB123wnco_1981_3.out

1981.10.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.67, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 13 August 2007

MatchOnMB12wco11pt67_AU1234wnco_1981_1.out
MatchOnMB12wco11pt67_AU1234wnco_1981_2.out
MatchOnMB12wco11pt67_AU1234wnco_1981_3.out
MatchOnMB12wco11pt67_AU1234wnco_1981_4.out
MatchOnMB12wco11pt67_AU1234wnco_1981_5.out
MatchOnMB12wco11pt67_AU1234wnco_1981_6.out

1981.10.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.67, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 13.65 – 14 August 2007

MatchOnMB12wco11pt67_AU12wco13pt65_1981_1.out
MatchOnMB12wco11pt67_AU12wco13pt65_1981_2.out
MatchOnMB12wco11pt67_AU12wco13pt65_1981_3.out
MatchOnMB12wco11pt67_AU12wco13pt65_1981_4.out

1981.10.4 Job 4 – passes 1 to 3, using residual records, blocking on meshblocks 1 to 3 on cancer file, with no cut-off – 14 August 2007

MatchOnRESp4MB123NCO_1981_1.out
MatchOnRESp4MB123NCO_1981_2.out
MatchOnRESp4MB123NCO_1981_3.out

1981.10.5 Job 5 – passes 1 to 4, using residual records, blocking on area units 1 to 4 on cancer file, with no cut-off – 14 August 2007

MatchOnRESp4AU1234NCO_1981_1.out
MatchOnRESp4AU1234NCO_1981_2.out
MatchOnRESp4AU1234NCO_1981_3.out
MatchOnRESp4AU1234NCO_1981_4.out

1981.10.6 Job 6 – passes 1 to 2, using residual records, blocking on meshblock 2 and area unit 2 on cancer file, with no cut-off – 16 August 2007

MatchOnRESp4MB2AU2NCO_1981_1.out
MatchOnRESp4MB2AU2NCO_1981_2.out

1981.10.7 Job 7 – passes 1 to 2, using residual records, blocking on meshblock 3 and area unit 3 on cancer file, with no cut-off – 5 September 2007

MatchOnRESp4MB3AU3NCO_1981_1.out
MatchOnRESp4MB3AU3NCO_1981_2.out

1981.11 List of PPV calculations for each job and pass

1981.11.1 Job 1 – passes 1 to 3, blocking on meshblocks 1 to 3 on cancer file, with no cut-off –10 August 2007

Passes123MB NCO_1981_lastppvPass1.out
Passes123MB NCO_1981_lastppvPass2.out
Passes123MB NCO_1981_lastppvPass3.out

1981.11.2 Job 2 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.67, passes 3 to 6, blocking on AU1 to AU4 (with sex), with no cut-off – 13 August 2007

Passes12MBwco11pt67_3to6AUnco_1981_lastppvPass1.out
Passes12MBwco11pt67_3to6AUnco_1981_lastppvPass2.out
Passes12MBwco11pt67_3to6AUnco_1981_lastppvPass3.out
Passes12MBwco11pt67_3to6AUnco_1981_lastppvPass4.out
Passes12MBwco11pt67_3to6AUnco_1981_lastppvPass5.out
Passes12MBwco11pt67_3to6AUnco_1981_lastppvPass6.out

1981.11.3 Job 3 – passes 1 to 2, blocking on meshblocks 1 to 2 on cancer file, with cut-off 11.67, passes 3 to 4, blocking on AU1 to AU2 (with sex), with cut-off 13.65 – 14 August 2007

Passes12MBwco11pt67_34AUwco13pt65_1981_lastppvPass3.out
Passes12MBwco11pt67_34AUwco13pt65_1981_lastppvPass4.out

Glossary

Array Where more than one value is available for the same variable, QualityStage allows the user to combine these values into an array to reduce the number of cross comparisons that must be made.

Block The files to be linked are divided into smaller subsets (blocks) which have some information in common. Records are only compared with others in the same block. Blocking reduces the number of comparisons that are made.

Blocking variable A variable used to break down large files into smaller subsets (blocks).

Clerical review The process by which comparison pairs (where it is unclear whether the pair should be linked or unlinked) are reviewed and assigned as linked or unlinked.

Cohort A group of people experiencing the same event in the same period of time. In this project a cohort for a census is the people with new cancer records in the period from the day after the census up to the day of the next census (inclusive).

Comparison pair Any possible comparison of a record from one file with a record from another file.

Cut-off (weight) The threshold weight above which pairs are accepted as links.

Duplicate or duplicate link(s) A record on one file that has two or more links with records on the other file.

False negative link A pair of records that is not linked when it is actually a match.

False positive link A pair of records that is linked when it is actually a non-match.

Identical links Pairs of the same records that were linked in different QualityStage jobs.

Job A series of passes carried out, in a set order, on two initial files, producing one set of linked records, two duplicate files, and two residual files.

Link A pair of records that is accepted as highly likely to apply to the same individual.

Linking variables Variables used to compare two records, including both blocking and matching variables.

Match A pair of records that apply to the same individual (ie true links).

Matching variables Variables used to compare two records that fall within the same block, to see how likely it is that two records belong to the same individual.

Non-link A comparison pair that is not accepted as being highly likely to apply to the same individual.

NZHIS New Zealand Health Information Service.

Pass The process of linking two files for a given specification of blocking variables, matching variables, *m* and *u* probabilities, and cut-off weight.

Probabilistic record linkage A record linkage methodology based on the relative likelihood that two records belong to the same person given a set of similarities/differences between the values of the linking variables on the two records.

Probability

- ***m*-probability (MPROB)** The probability that a matching variable agrees, given that the comparison pair is a match. This probability generally reflects the accuracy of the recorded data.

- ***u*-probability (UPROB)** The probability that a matching variable agrees, given that the comparison pair is a non-match. This probability is generally determined by the likelihood of both records having the same value due to chance, and reflects the incidence of common data values.

Positive predictive Value (PPV) The proportion of linked records that are true positives (ie matches).

Residuals The remaining unlinked records after a job has been run.

True negative A pair of records that is not linked when it is actually a non-match.

True positive A pair of records that is linked when it is a match.

Weight

- **Agreement weight** The value assigned for agreement on a given matching variable. This value is a positive number, calculated from the *m* and *u* probabilities for that variable according to the following formula:

  $[\ln(m/u)/\ln(2)]$

- **Disagreement weight** The value assigned for disagreement on a given matching variable. This value is a negative number, calculated from the *m* and *u* probabilities for that variable according to the following formula:

  $[\ln((1 – m)/1 - u)/\ln(2)]$

- **Total weight (or weight)** The sum of the agreement and disagreement weights for each matching variable in a comparison pair of records.

References

Blakely T and Salmond C (2002). "Probabilistic record linkage and a method to calculate the positive predictive value", in *International Journal of Epidemiology*, 31, 1246–1252.

Smith A (2005). "CancerTrends: Technical Report & Feasibility Assessment", Lotus Notes database, Statistics New Zealand, Wellington.