CRIME AND JUSTICE
Bulletin

NSW Bureau of Crime Statistics and Research

Contemporary Issues in Crime and Justice Number 95

July 2006

Matching Court Records to Measure Reoffending

Jiuzhao Hua and Jacqueline Fitzgerald

In 2001, the NSW Bureau of Crime Statistics and Research developed a reoffending database. This database links individuals' criminal court appearance records over time and so enables the measurement of recidivism in NSW. This bulletin describes the results of a validation technique we applied to test the accuracy of the matching processes that underpin the reoffending database. We first describe the development of the reoffending database and the deterministic matching criteria that are its foundation. We then describe the validation technique of applying those deterministic matching criteria to a dataset of individuals with known identities and measuring the number of false positives and false negatives that resulted. The validation results suggest that the error rates in the reoffending database are likely to be acceptably low. Finally we discuss the testing of an additional matching criterion involving residential postcode and recommend its adoption.

INTRODUCTION

For crime reduction efforts to be successful, it is critically important to understand recidivism and the characteristics of recidivist offenders. The importance of monitoring reoffending is underscored by the fact that a small minority of offenders account for a large proportion of offences (see, for example, Coumarelos 1994, Salmelainen 1995, Baker 1998). To facilitate such monitoring, the NSW Bureau of Crime Statistics and Research (BOCSAR) has developed a Reoffending Database (ROD) which tracks the frequency with which individuals appear in NSW criminal courts (see Weatherburn, Lind & Hua 2003 for a description of the development of ROD). Other Australian jurisdictions have also built databases to track individuals through the criminal justice system (see, for instance, a description of the Western Australian model in Ferrante 1993).

Since its development in 2001, ROD has proven to be extremely valuable in determining rates of reoffending, providing evidence to inform program development and measuring the impact of criminal justice interventions on offending. Some of the research and policy projects ROD has been used for include:

- Measuring the proportion and characteristics of juvenile offenders who go on to appear in the adult court system
- Measuring the impact of increased drink-driving penalties on recidivism
- Measuring reoffending among parolees
- Estimating the number of persons eligible for the Compulsory Drug Treatment Correctional Centre to be trialled in 2006
- Estimating the number of persons eligible for the Community Conferencing for Young Adults program

ROD is built by linking the court records of individuals. The purpose of this bulletin is to describe the results of a validation test applied to the matching criteria used in ROD. First we discuss the matching criteria. Then we describe the strategy we used to test the accuracy of the ROD matching. Finally we discuss an enhancement we have made to the matching criteria as a result of the validation process.

THE ROD MATCHING PROCESS

ROD contains records of all finalised criminal court appearances in the Children's, Local, District and Supreme Courts of New South Wales since 1994. Children's Court data was provided to the Bureau by the NSW Department of Juvenile Justice. Data for the other jurisdictions were collected by BOCSAR. The database contains about 1,454,000 court appearance records for the period January 1994 to September 2005, which the ROD matching process indicates were generated by 690,000 distinct people.

The process of determining whether two (or more) court appearance records belong to the same or different people is far from simple or straightforward as there is no courts-generated common identifier. There are also many problems with the quality of information in court appearance records, including missing data, typing errors, spelling mistakes, transposition errors, abbreviations and misinterpretations of handwriting. Even the police-generated Central Names Index (CNI) number, intended to identify unique individuals, is of limited assistance because 34 per cent of the court records in ROD do not include a CNI and some individuals have more than one CNI.[1]

To overcome such deficiencies in the data, a data matching system is needed. Two commonly used data matching techniques are probabilistic matching, as is used in the Western Australian database (Ferrante 1993), and deterministic or rules based matching (both techniques are described in Christen and Goiser 2005). ROD employs a purpose-built deterministic system to conduct its matching.

The ROD matching criteria were developed with particular consideration for the limits and strengths of the personal identifying information collected in NSW Criminal Courts. During the development of ROD, the ability to match individuals using various criteria was tested. The accuracy and appropriateness of the criteria were tested by extensive visual inspection of sampled results. Five final criteria were decided upon as it was considered that these generated the highest number of valid matches without producing an unacceptable number of obvious errors. (Of course, at the time of development our only test of accuracy was visual inspection of the matches made, and not made, from the flawed input data. There was no independent means available to us to test whether two people were indeed the same person or not outside of the potentially flawed data in the court records.)

The matching procedures developed for ROD are listed below. ROD decides whether the defendants in two (or more) court records are likely to be the same individual by comparing their personal identifying particulars against the five sets of matching criteria. If two court appearance records match according to at least one of these sets of criteria, they are deemed to involve the same person. If not, they are deemed to involve distinct persons.

To be matched under the ROD matching criteria, two records must have the same:

1. Surname, First name and Date of birth (DOB); or
2. Surname, First name, Middle name and two components of the DOB[2]; or
3. CNI and DOB; or
4. CNI, Surname and two components of the DOB; or
5. CNI, First name and two components of the DOB

When applied to the court record data, some of the five criteria are responsible for more matches than others. This can be demonstrated by applying the criteria in isolation. When just the two name-based criteria (1 & 2) are run across the court record data, they identify 30 per cent of records as matches. When just the three CNI-based criteria (3, 4 & 5) are applied (without the name criteria) a match rate of only 14 per cent is returned. When the five criteria are applied together they identify 34 per cent of records to be matches. Thus, the name-based criteria contribute the most matches. Note, gender is not used in any of the criteria.[3]

When comparing names, dates of birth and CNIs, ROD uses certain variations of the input to try and obtain a match. For instance, two names are considered to be the same if they match:

- by soundex[4], that is they have similar sounds but different spellings. (e.g. 'Steven' and 'Stephen' are matched by soundex)
- with one letter dropped. (e.g. dropping the letter 'P' matches 'Thomson' with 'Thompson')
- with the surname and first name reversed. (e.g. 'Sebastian James' is matched to 'James Sebastian')
- using a common abbreviated form.[5] (e.g. 'Benjamin' is matched to 'Ben')

Two dates of birth are considered to be the same if they match after swapping either:

- the day and month of birth (e.g. 9.02.1976 is matched to 2.09.1976); or
- the last digits of the day and month (e.g. 19.02.1976 matches 12.09.1976).

One potential limitation of deterministic matching is that it can require the comparison of an untenably large number of records. Comparing every court appearance record in our database with every other court appearance record in order to determine whether they were the same would require an impossible amount of computer processing. To limit the number of comparisons necessary, ROD uses a process of sorting. In this process, records are variously sorted according to different fields so that each record only needs to be compared to those with similar values.

POSSIBLE MATCHING ERRORS

This paper is interested in how well the matching criteria reported above perform in joining records belonging to the same individual and in keeping separate records for different individuals. There are two potential types of errors that can occur in linking court records to individuals:

- ROD can fail to link two court appearance records that actually belong to the same person. This is called a *false negative*; or
- ROD can link court appearance records that actually relate to two distinct persons. This is called a *false positive*.

It is possible to influence the occurrence of these errors by modifying the matching criteria listed above. We could reduce the false negative rate, for example, by allowing a match based on just two components of the date of birth instead of the requirement in matching criteria (1) that the two records have precisely the same date of birth. There is an inverse relationship, however, between the likelihood of the two matching errors. Allowing less precise matches reduces the number of false negatives but increases the false positive rate. Requiring more precise matches, on the other hand, reduces the false positive rate but increases the false negative rate.

Until recently, it was impossible to estimate the number of false positives and false negatives made in the ROD matching process because we had no independent way of ascertaining which court records belonged to the same individual and which did not. Validation consisted of little more than visual inspection of records matched by ROD to see whether they appeared to be correct matches or not. (Appendix 1 shows examples of court records which ROD considers involve the same person, but which may in fact involve different people.)

One possible way to validate data matching techniques is to engage in a clerical review of a subset of records. In such a process we could compare criminal histories generated by ROD with those from another source, such as the criminal histories maintained by the NSW Police. The process of clerical review, however, would be time consuming and require the provision of external data. Realistically, without significant resources, it could only be performed on, at most, several hundred records, a tiny subset of the estimated 690,000 distinct person records in ROD.

An entirely different validation method was used for the current project. Here we attempted to assess the reliability of the matching criteria through the use of a large group of distinct individuals whose true identities were known. This enabled us to determine whether, and to what extent, the matching criteria correctly identify distinct individuals. In the next section of this bulletin we report the results of our analyses designed to use data drawn from a set of birth records to assess the accuracy of the matching criteria employed in ROD.

MATCHING VALIDATION PROCESS

There were two stages to the validation of the ROD matching criteria. In the first stage, the ROD matching criteria were applied to a group of distinct individuals to determine the frequency with which our matching programs generate *false positives*. In the second stage, we created a 'virtual' ROD to estimate the frequency of *false negatives*. Note that, because our sample had no CNI equivalent, our validation testing was restricted to errors arising through matching on name and DOB. This means we were limited to testing the first two sets of ROD matching criteria described above, namely: (1) surname, first name, DOB and (2) surname, first name, middle name and two components of DOB. Fortunately, these two criteria are the most important of the five in terms of the proportion of matches they contribute.

THE 1984 BIRTH COHORT

The dataset of distinct individuals used for the validation consisted of all persons born in NSW in 1984. In that year there were 83,042 births registered with the NSW Registry of Births, Deaths and Marriages.[6] All birth records have a surname, first name and DOB. Eighty six per cent of records have a middle name.[7] Other recorded information on the cohort includes name, age and residential suburb of the newborn's parents and the names of any siblings. Each recorded birth also has a registration number.

Cleaning the birth cohort

If the 1984 birth cohort dataset was to provide a reliable basis on which to validate ROD, it could not contain duplicate records belonging to the same individual. However, an individual could appear in the birth cohort twice if his or her birth was mistakenly registered twice. To identify any duplicates, the birth cohort dataset was searched for persons with the same name and date of birth. The resulting matches were then manually checked to determine whether or not they were the same person. By using the extra information available on birth records, such as address and the names and ages of the newborn's parents, it was possible to determine if two records belonged to the same person. Out of the 83,042 records in the birth cohort, 366 duplicates were identified and removed, leaving 82,676 unique persons.[8]

FALSE POSITIVES IN THE POPULATION

Once all duplicates were removed from the 1984 birth cohort the number of matching errors was estimated. To estimate the frequency of *false positives*, the ROD matching criteria were applied to each member of the birth cohort. As each record in the cohort represents a unique person, any match that occurs must be false. There is no opportunity to *miss* a match in this part of the validation, as there should not be any genuine matches. The error rate or *false positive* rate[9] was determined as follows:

$$\text{False positive rate} = \frac{\text{Number of unique people falsely matched}}{\text{Number of distinct individuals}}$$

$$= \frac{\text{Number of false positives x 2}}{\text{Number of distinct individuals}}$$

Table 1: Examples of distinct individuals from the 1984 birth cohort incorrectly matched by the ROD matching criteria

Person	*Surname*	*First Name*	*Middle Name*	*DOB*	*Mother's Age*	*Mother's Name*
1	Wilson	Daniel	John	24/06/1984	26	Diane Lee Wilson
2	Wilson	Daniel	Leonard	24/06/1984	32	Sue Margaret Wilson
3	Taylor	Matthew	James	16/12/1984	22	Sharon Rachael Taylor
4	Taylor	Matthew	James	16/12/1984	25	Gwen Yvonne Taylor
5	Smith	Jessica	Emma	18/08/1984	28	Raelene Sarah Smith
6	Smith	Jessica	Emma	18/03/1984	24	Patricia Suzanne Green

Note: Not real individuals

Subsequent false positive error rates are calculated in a similar fashion. The results of the matching process are shown below:

Number of distinct persons:	82,676
Number of false positives:	125

False positive rate:
(125 x 2) / 82,676 x 100 = 0.30%

There were 125 false positives, giving a false positive rate of 0.30 per cent. In other words, for every 667 people in the population, one would have personal information that is similar enough to another distinct person that they would be considered to be the same person using the ROD matching criteria.

Two people could be falsely matched if they had either the same surname, first name and date of birth; *or* if they had the same surname, first name and middle name *and* were born either in the same month *or* on the same day of different months (this would satisfy the two components of the DOB requirement as all were born in 1984). Table 1 shows some examples of the types of false positives made within the 1984 birth cohort. To avoid the identification of real people, the names and dates of birth have been altered. It is apparent that these people are not the same individuals because their mothers' names and ages are different.

The first two persons in Table 1 match on surname, first name and DOB even though they have different middle names. The third and the fourth persons match on surname, first name, middle name and DOB. It is quite rare to have two people with exactly the same name born on the same day. This generally only happens when the names are very popular. The last two persons match on surname, first name, middle name and two components of the DOB and yet they are distinct persons.

It is very difficult to avoid matching two people who have the same or very similar names and the same date of birth. In the 1984 birth cohort it is possible to tell whether two people are distinct by using information about their mothers, however, this information is not available in court records. Nevertheless it should be noted that, while it is inevitable that some false positives will occur in the ROD matching process, the checks conducted here suggest that it is quite uncommon.

MATCHING ERRORS IN A 'VIRTUAL' ROD

The process described above does not tell us what the false negative rate is in ROD as the 1984 birth cohort is comprised of distinct individuals so there are no matches to be missed. In addition, ROD is only concerned with linking the records of the subset of the population who appear in court. People who have never appeared in court have no records in ROD, while some people appear once and others repeatedly. To model this for the next stage of the validation, we built a 'virtual' ROD by estimating the number of people in the 1984 birth cohort with a court appearance and the frequency distribution of those appearances. We then deliberately added some errors (by approximating those that exist in actual court records) and calculated the frequency of false negatives in 'virtual' ROD that occurred after running our ROD matching programs.

Building a 'virtual' ROD

We began building the 'virtual' ROD by estimating the number of court appearances that the 1984 birth cohort was likely to have had between 1994 and 2004. Figure 1 shows the frequency distribution of court appearances for all people who actually appeared in court between 1994 and 2004.

It can be seen that, among those who appeared in court between 1994 and 2004, 63 per cent had only one court appearance, 16 per cent had two court appearances and seven per cent had three court appearances. The remaining 14 per cent had four or more court appearances. The next step was to use the data in Figure 1 to derive what is known as a hazard function in order to approximate reoffending.[10] Figure 2 shows the hazard function derived from Figure 1.

Figure 2 indicates that, among those who had at least one court appearance, 37 per cent went on to have a second court appearance. Of those who had at least two court appearances, 56 per cent went on to have another court appearance. Beyond a person's seventh court appearance the hazard function is quite stable, with about an 80 per cent chance of reappearing. To construct the 'virtual'

Figure 1: Unique persons appearing in NSW Criminal Courts between 1994 and 2004 by number of court appearances in the period

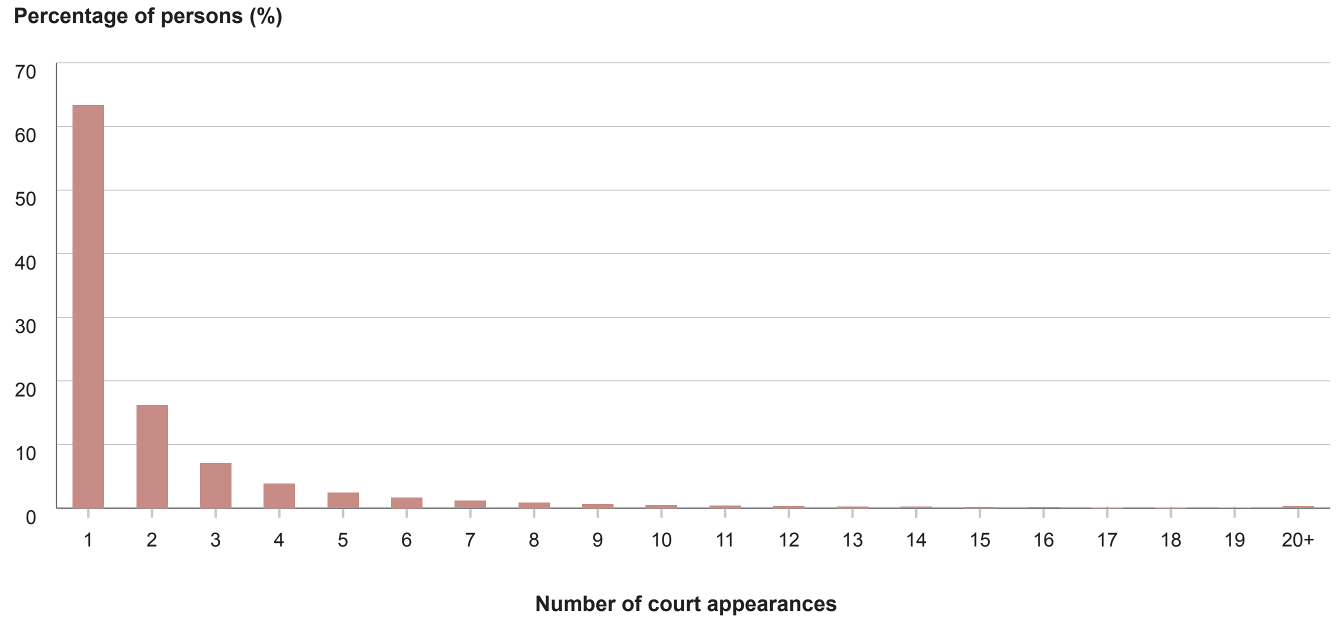
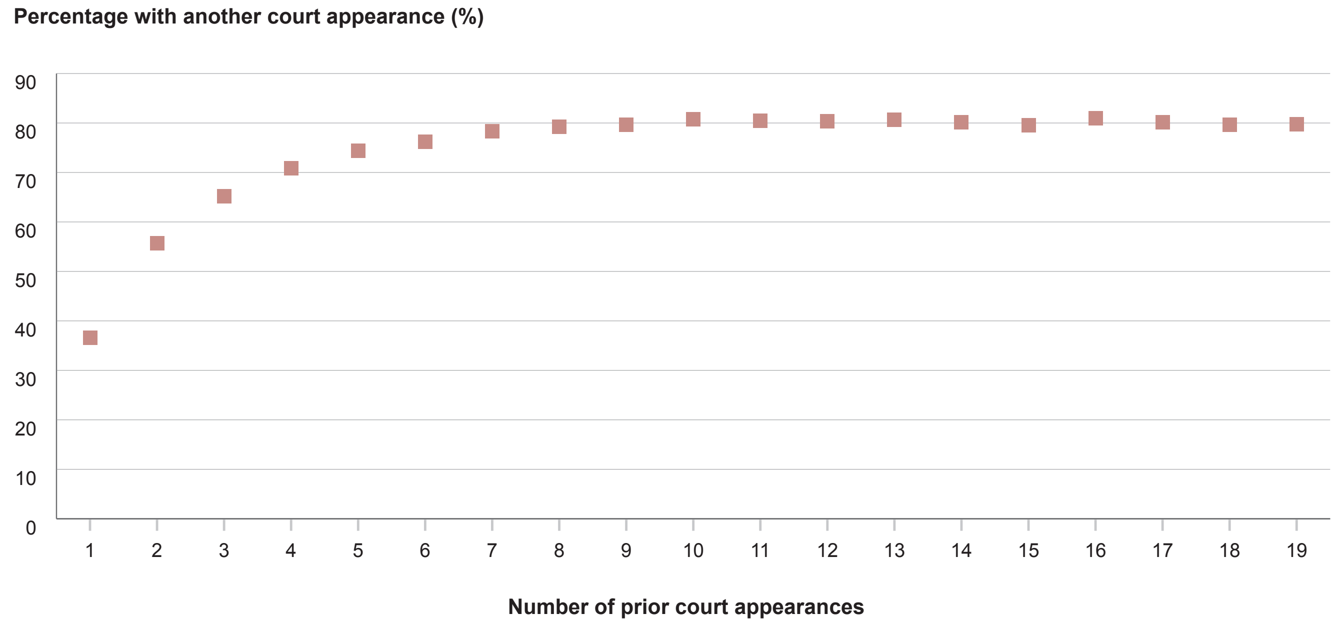


Figure 2: Hazard function showing probability of a person reappearing in court as a function of the number of prior appearances



ROD, we first estimated the proportion of the cohort with at least one court appearance. We then created duplicate ('reappearance') records from our birth cohort in proportions mirroring those shown in Figure 2.

Evidence indicates that about 30 per cent of people in our cohort would have had at least one court appearance between 1994 and 2004.[11] We therefore randomly selected 30 per cent of the 1984 birth cohort to represent the proportion of the population with a court appearance.

From this group, 36 per cent were then randomly selected to represent the portion of the sample with at least two court appearances. From the group with at least two court appearances, 57 per cent were randomly selected to represent those with three or more court appearances. This process was repeated using the hazard function from Figure 2, until the proportion of the cohort with 19 reappearances had been estimated.

Since the actual court records used to build ROD include misspellings, mistypings and incomplete data, the next stage in building the 'virtual' ROD was to incorporate errors in the data similar to those that exist in the actual court records. Unfortunately, there is no way to know the true rate of errors in the actual data. We assumed an error rate of five per cent, which means that one in twenty names and DOBs contain an incorrect character. Given the prevalence of electronic data transfer in current records, we consider that this is a reasonably high proportion of errors and probably exceeds the reality. Errors were then assigned to each court appearance record according to the following assumptions:

- There is a five per cent chance that one character in the surname will be mistyped, (e.g., SMITH would be typed as SMITS). Each character in the name had an equal chance of being mistyped
- There is a five per cent chance that one character in the first name will be mistyped
- There is a five per cent chance that one character in the middle name will be mistyped
- There is a five per cent chance that either the day, month or year in the date of birth will be mistyped, e.g. 19/01/1984 would be typed as 16/01/1984
- The 'errors' in the surname, first name, middle name and date of birth happen independently of each other. Thus, a record could have errors in both the surname and the DOB

The assumed errors were randomly applied to the court appearance records generated by the hazard function for the 1984 birth cohort. So, five per cent of the court appearances we estimated for the 1984 birth cohort had the surnames altered by one character, five per cent had the first name altered by one character and so on.

Matching results

The ROD matching criteria were then applied to the 'virtual' ROD database to see how many false positives there were and how many matches were missed using the regular matching criteria. The number of false negatives is the number of duplicates that the criteria failed to identify. The false negative rate is determined by calculating from the formula below.

The outcome of the matching process was:

Number of distinct persons:	24,916
Number of court appearances generated:	52,944
Number of false negatives:	1,740
False negative error rate:	6.2%
Number of false positives:	15
False positive error rate[12]:	0.057%

Given the uses to which ROD is being put, there is probably more harm associated in *false positives* than with *false negatives*. An excessive number of false positives would mean the criminal histories of multiple people would be attributed to single individuals. As a result, the database would overestimate the number of recidivist offenders and the extent of their recidivism. On the other hand, by erring towards not matching individuals who might actually be the same person, we end up with a conservative measure of recidivism. We can then claim with certainty that the rate of reoffending is at least that shown by ROD.

Given this, it is reassuring to note that false positives are very uncommon when linking people by name and DOB. Out of 52,944 court appearances by 24,916 persons, there were only 15 cases of distinct people being matched.[13] The more common, but less worrisome, type of error is the failure to connect court appearances for the same people, that is, the false negative rate. The rate of this type of error is 6.2 per cent, which means that, out of every 16 real matches, one is not identified by our criteria.

IMPROVING THE ACCURACY OF ROD MATCHING

MATCHING ON RESIDENTIAL POSTCODE

The process described above shows the rate of *false positives* to be very low in comparison to the rate of *false negatives.* Our next step was to see whether the number of *false negatives* could be reduced by trying an additional matching criterion. Note that the rate of *false positives* is so low, a modest increase could be tolerated if it resulted in a reduction in the rate of *false negatives.*

Postcode of residence offers another potential individual identifier as it is recorded on each court appearance record in ROD. Residential postcode, however, is not an ideal identifier because it changes when people move residence. Since an individual can legitimately have different postcodes at different court appearances, the postcode matching criterion is only useful in matching people who have not changed postcodes between court appearances. When the original ROD matching criteria were developed, postcode was not included because of both its changeable nature and because at that time we had no way of independently measuring the validity of matching on postcode.

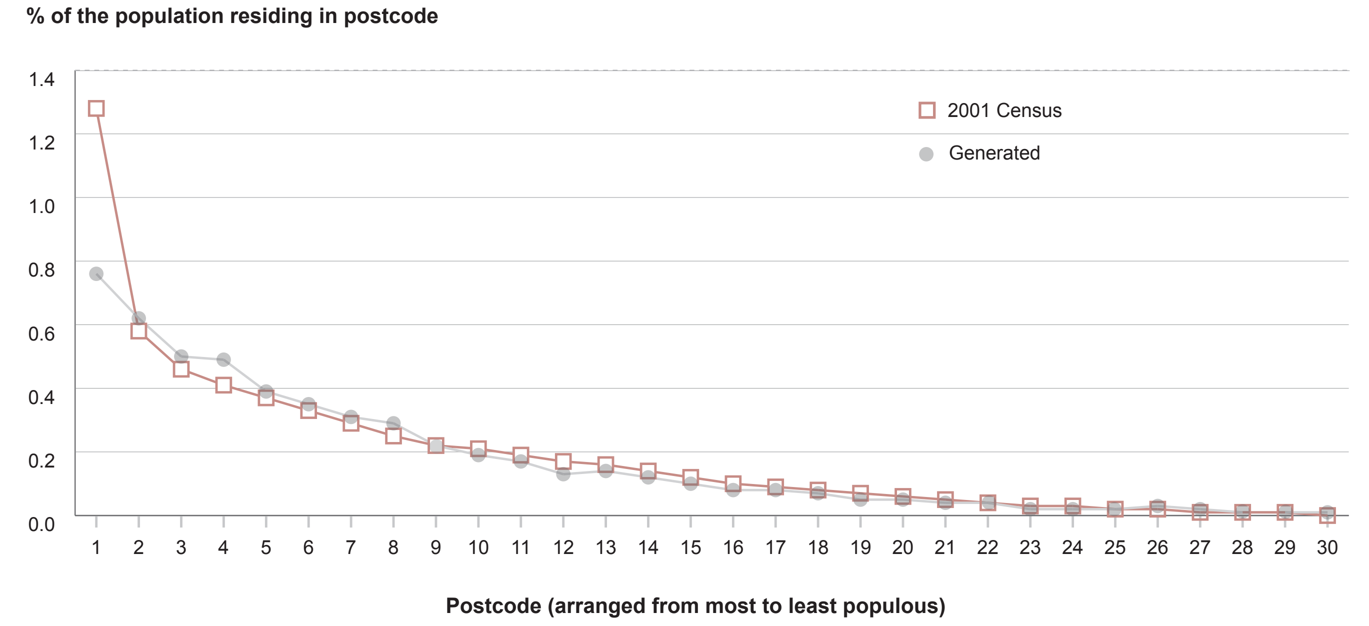
'Virtual' ROD allowed us to check whether matching on postcode would be useful and, if so, how it would affect the error rate. The criterion we were interested in testing, therefore, was:

- Surname, first name, two components of the DOB, postcode

In order to estimate the impact and reliability of this new criterion, it was necessary to include a degree of

$$\text{False negative rate} = \frac{\text{Number of false negatives}}{\text{Number of court appearance records - Number of distinct persons}}$$

Figure 3: Percentage of persons born in 1984 by postcode: 2001 Census and generated



Note: To simplify the figure, only one in 20 postcodes are shown. These postcodes are represented by sequential integers.

geographic mobility into our 'virtual' ROD. To do this, a randomly generated postcode was added for each individual. The assignment of postcodes was proportional to the number of people resident in each postcode in NSW. Geographic mobility was simulated, by assuming that:

- There is a 50 per cent chance that a person will change his or her postcode after each court appearance;[14] and
- When a person moves, the chance that they will move to a particular postcode is proportional to the size of the population of that postcode.

The proportion of people assumed to be living in each postcode was determined from the actual population distribution among NSW postcodes. Figure 3 shows two data series.[15] The first series shows the percentage of the population residing in each NSW postcode according to the 2001 Census. This series appears to be well-fitted by an exponential distribution. We used this fact to generate the postcodes for our simulation of geographic mobility. The second series in Figure 3 estimates the residential postcode of persons born in 1984 derived from the 2001 Census and our simulation. The two series align very closely (with the one exception of the most populous postcode).

MATCHING RESULTS

After a postcode was allocated to each court appearance record according to the assumptions described above, the ROD matching criteria, including the postcode criterion, were applied to the 'virtual' ROD model. The results are shown below.

Number of distinct persons:	24,916
Number of court appearances generated:	52,944
Number of false negatives:	1,234
False negative error rate:	4.4%
Number of false positives:	15
False positive error rate:	0.057%

The number of court appearances generated and the number of distinct persons in this matching process are the same as in the previous analysis because we used the same base dataset. Inclusion of the postcode criterion, however, gave an additional 506 real matches with no extra false positives. The percentage of false negatives reduced from 6.2 per cent to 4.4 per cent. This finding strongly supports the inclusion of postcode as one of the matching criteria.

It is surprising that our postcode simulation did not result in a single additional false positive. The simulation is based on applying conditions (such as error rate, multiple court appearances and changing postcodes) to randomly selected subsets of the sample. This means that error rates will vary somewhat from one simulation to the next depending on which records happen to be randomly selected and varied. This simulation may have, by chance, shown an unusually low number of false positives. It is reasonable to assume that when the postcode criterion is applied to the complete ROD database some false positives will arise. The results above, however, suggest that the number will be small.

SUMMARY AND CONCLUSION

ROD is an important research and policy tool but its usefulness depends entirely on its reliability. This bulletin describes an attempt to validate the accuracy of the ROD matching process using an independent cohort of distinct individuals. A 'virtual' ROD database was built from the cohort of people born in NSW in 1984 by first estimating the likely number of court appearances among the group and then incorporating a fixed proportion of data errors. The ROD matching criteria for names and DOB were then applied to the 'virtual' database to see how often they resulted in errors.

The error rates were low. The rate of false positives was estimated at 0.057 per cent. In other words, for every 10,000 people in the database, only three were incorrectly matched to another person. The rate of false negatives was estimated at 6.2 per cent. This means that, for every 16 actual matches that exist in the data, one is not identified in our criteria. In reality, this is unlikely to be the actual rate of false negatives in ROD; the actual ROD error rate could be lower due to the possibility of matching using CNI, alternatively the ROD data could contain errors not taken into account here which would give a higher error rate. Although the estimated rate of false negatives is considerably higher than the estimated rate of false positives, in most applications of ROD it is probably better to miss a match than to mistakenly make one.

The ROD simulation made it possible to test the viability of an additional matching criterion involving postcode of residence. This additional criterion was found to reduce the rate of false negatives to 4.4 per cent, without increasing the number of false positives.

There are, however, some weaknesses in our simulation models which should be considered:

- In the community, the likelihood of offending varies for different subgroups. For instance, it would be expected that males and Indigenous people in the 1984 birth cohort would have higher rates of contact with the court system. Such differences were not incorporated in the model.[16]
- Because our test group were all born in the same year, the study did not measure the possibility of matching distinct people, whose personal identifying information is the same with the exception of birth year.
- The models did not control for significantly corrupted names. In some cases a whole component of a person's name might be entered incorrectly, for instance Teddy might be recorded as Gerry. These kinds of errors will result in false negatives, but were not incorporated into these analyses.
- The models are based on assumptions, some of which cannot be verified. For instance we assumed that one in twenty surnames include an incorrectly typed character. We have no way to test how accurate this assumption is.
- The models only test two of the existing five ROD matching criteria (and provide support for the inclusion of a sixth). Three of the existing criteria based on CNI remain untested. (It should be noted that the two matching criteria based on name are responsible for 88 per cent of matches in ROD; only 12 per cent are made on the CNI criteria alone.)

Despite these limitations, our validation process provides evidence that, where it has been tested, the ROD matching process is highly reliable. Consequently, estimates about reoffending generated from ROD are likely to be sufficiently accurate for statistical and research purposes.

ACKNOWLEDGEMENTS

The authors extend their gratitude to the NSW Registry of Births Deaths and Marriages for making the 1984 birth cohort data available. Without this data the validation of ROD would not have been possible. Thanks also to Dr Don Weatherburn, Tracy Painting, Steve Moffat and Anna Ferrante for their valuable critical feedback on earlier versions of this paper.

NOTES

1. CNI is a unique person identifier assigned to suspects by NSW Police. In the 1990s, CNI was often missing from court documents lodged by the police in NSW; this is no longer the case, mainly due to the introduction of electronic court lodgements. Children's Court appearance records prior to 2006 did not include CNI. It is still the case, however, that defendants brought to court by agencies other than police, such as the Australian Tax Office, Local Councils and the RSPCA, do not have a CNI. Another problem with CNI is that some individuals have more than one. This occurs when police fail to recognise that an offender or suspect already exists in their system and assign the person a new CNI. Fingerprinting is the most accurate way to determine whether a person is already in the police system. However, many offenders/suspects are not fingerprinted, especially those processed in the field (or away from the police station) which is increasingly common. The propensity of offenders to use aliases and give false personal details also contributes to this problem.

2. Two components of the date of birth give a match if two out of the day, month and year are identical. For instance, 12/08/1984 is matched to 28/08/1984 by the two components of the DOB rule. Regardless of the other components, two DOBs are not matched if they contain birth years that are more than ten years apart.

3. Gender is not used in ROD. This is because the ROD matching technique is based on searching the input data for evidence that records belong to the same people rather than finding evidence that they do not. Since gender only has two values, it does not offer any real confirmation that two people are the same. However, if two records matched on one of the sensitive personal information items, such as name or CNI, but not on gender, it would be more likely that the gender was wrongly recorded than that the records involved different people

4. Soundex codes work by converting words to codes. Letters in specified groups are given the same value. For example the letters 'y' and 'ie' could be placed in the same group, given the same code and therefore be regarded as identical. We developed our own soundex code for

ROD, expanding on the SAS Soundex matching options.

5. The common abbreviated forms were compiled by BOCSAR and are mostly limited to common variations of Western first names. Unfortunately, common variations of names from other cultures are not well represented.

6. The NSW Bureau of Crime Statistics and Research obtained the details of the 1984 birth cohort from NSW Registry of Births, Deaths and Marriages.

7. Only 59% of court appearance records in ROD have a middle name recorded. The discrepancy in the number of registered births with a middle name and the number of persons appearing in court with a middle name gives some indication of the imprecision of the court data.

8. This means that approximately 0.4 per cent of records in the original cohort of births registered in 1984 were duplicates.

9. The authors are aware that other analysts, for instance Gu et al. 2003 and Christen & Goiser 2005, recommend different formulae and terms to describe matching errors. These were not used in this paper, as they all require the calculation of the actual number of genuine matches, genuine non-matches, false positives and false negatives. Calculating these inputs would require a pairwise comparison of all records and would only be possible with a small dataset. Our test data contains more than 50,000 records, which would require 50,000 x 50,000 = 2,500,000,000 comparisons. BOCSAR does not have computer hardware capable of comparing this number of records.

10. In the present context, the hazard function is the probability of reappearing in court n times, given n-1 prior appearances.

11. The figure of 30 per cent is derived from other studies estimating the proportion of the population with a conviction. See, for instance, Tarling 1993.

12. A key feature of 'Virtual' ROD is that some individuals appear more than once as they have been estimated to have more than one court appearance. For this reason, in calculating this false positive rate, the appropriate denominator was the total number of court appearances generated (not the number of distinct individuals).

13. It is worth noting that the false positive rate from 'virtual' ROD is much lower than the false positive rate we saw for the entire cohort (0.30%). This is because each person in the cohort has only a 30 per cent chance of appearing before the courts. Therefore, the chance that the two distinct persons who were previously matched both have contact with the court is 0.3 x 0.3, which is only nine per cent.

14. Note that, although we have assumed that 50 per cent of people move between court appearances, we do not know the true figure. It is likely that 50 per cent would actually overstate the mobility of people between court appearances. Between 1996 and 2001, 42 per cent of Australians aged five or over moved residence (ABS 2003). Thirty two per cent of these people moved within the same Statistical Local Area (which would sometimes, but not always, be in the same postcode). Thus, a sizeable proportion of the population do not move in five years and many who do, stay in the same neighbourhood. However, because people who appear in court are likely to be more transient than the rest of the community, we have estimated a higher mobility rate.

15. In order for the two series plotted on this figure to be seen, only one in every twenty postcodes is shown. For purposes of illustration it is not necessary to show the actual postcodes, so in Figure 3 postcodes are represented by sequential integers.

16. Males account for about five out of every six court convictions. The high rate of male offending could have been incorporated into the model by attributing more court appearances to males from the 1984 birth cohort. This would not change the distribution of family names or DOBs; however, there would be more repetition among first and middle names due to the higher prevalence of records from males. This would not be expected to alter the false negative rate, but would probably give an increase in false positives.

REFERENCES

Australian Bureau of Statistics 2003, *Census of Population and Housing: Population Growth and Distribution Australia 2001,* cat. no. 2035.0, ABS, Canberra.

Baker, J 1998, *Juveniles in crime - Part 1: Participation rates & risk factors,* NSW Bureau of Crime Statistics and Research, Sydney.

Christen, P & Goiser, K 2005, 'Assessing Duplication and Data Linkage Quality: What to Measure?', *Proceedings of the fourth Australasian Data Mining Conference,* Sydney, December 2005, viewed 16 June 2006, <http://datamining.anu.edu.au/linkage.html>.

Coumarelos, C 1994, *Juvenile offending: Predicting persistence and determining the cost-effectiveness of interventions,* NSW Bureau of Crime Statistics and Research, Sydney

Ferrante, A 1993, 'Developing an Offender-Based Tracking System: The Western Australia INOIS Project', *Australian and New Zealand Journal of Criminology,* 26, pp. 232-250.

Gu L, Baxter R, Vickers D and Rainsford C 2003 *Record Linkage: Current Practice and Future Directions,* CSIRO Mathematical and Information Sciences Technical Report 03/83, viewed 16 June 2006, <http://www.act.cmis.csiro.au/rohanb/PAPERS/record_linkage.pdf>.

Salmelainen P 1995, *The correlates of offending frequency: A study of juvenile theft offenders in detention,* NSW Bureau of Crime Statistics and Research, Sydney.

Tarling R 1993, *Analysing Offending: Data, Models and Interpretations,* HMSO, London.

Weatherburn, D, Lind, B & Hua, J 2003, Contact with the New South Wales court *and prison systems: The influence of age, Indigenous status and gender,* Crime and Justice Bulletin 78, NSW Bureau of Crime Statistics and Research, Sydney.

APPENDIX 1

Appendix 1: Examples of court appearances considered by ROD to involve the same offender

Court Appearance	*Surname*	*First Name*	*Middle Name*	*DOB*	*CNI*
1	Williams	Michael	Luke	25/03/1988	4837355
2	Williams	Mike	Luke	25/03/1978	6128074
3	Pappas	George	Alexander	16/01/1973	2661248
4	Pappas	George		11/06/1973	3562189
5	Jorge	Alyson	Judy	19/01/1969	
6	Jorge	Alison	Judy	19/10/1969	1578897
7	Porter	Genevieve	Grace	18/01/1974	1433062
8	Grace	Genevieve	Porter	18/01/1974	
9	Chan	Li	Mei	20/01/1980	2855769
10	Chan	Li Mei		20/01/1980	8759442
11	Le Breton	Paul	Denis	06/11/1947	1145569
12	Breton	Paul	Denis	06/11/1946	

Note: Not real individuals

Other titles in this series

No.94	Victims of Abduction: Patterns and Case Studies
No.93	How much crime does prison stop? The incapacitation effect of prison on burglary
No.92	The attrition of sexual offences from the New South Wales criminal justice system,
No.91	Risk of re-offending among parolees
No.90	Long-term trends in property and violent crime in NSW: 1990-2004
No.89	Trends and patterns in domestic violence
No.88	Early-phase predictors of subsequent program compliance and offending among NSW Adult Drug Court participants
No.87	Driving under the influence of cannabis: The problem and potential countermeasures
No.86	The transition from juvenile to adult criminal careers
No.85	What caused the recent drop in property crime?
No.84	The deterrent effect of capital punishment: A review of the research evidence
No.83	Evaluation of the Bail Amendment (Repeat Offenders) Act 2002
No.82	Long-term trends in trial case processing in NSW
No.81	Sentencing drink-drivers: The use of dismissals and conditional discharges
No.80	Public perceptions of crime trends in New South Wales and Western Australia
No.79	The impact of heroin dependence on long-term robbery trends
No.78	Contact with the New South Wales court and prison systems: The influence of age, Indigenous status and gender
No.77	Sentencing high-range PCA drink-drivers in NSW
No.76	The New South Wales Criminal Justice System Simulation Model: Further Developments
No.75	Driving under the influence of cannabis in a New South Wales rural area
No.74	Unemployment duration, schooling and property crime
No.73	The impact of abolishing short prison sentences
No.72	Drug use monitoring of police detainees in New South Wales: The first two years
No.71	What lies behind the growth in fraud?
No.70	Recent trends in recorded crime and police activity in Cabramatta
No.69	Reducing Juvenile Crime: Conferencing versus Court
No.68	Absconding on bail
No.67	Crime increases in perspective: The regional dispersion of crime in NSW, 2001
No.66	Hung juries and aborted trials: An analysis of their prevalence, predictors and effects
No.65	Multiple drug use among police detainees
No.64	Law enforcement’s Role in a Harm Reduction Regime
No.63	Do targeted arrests reduce crime?
No.62	Trends in sentencing in the New South Wales Criminal Courts: 1999-2000
No.61	Preventing Corruption in Drug Law Enforcement

Other titles in this series

No.60	Reducing Cannabis Consumption
No.59	The Australian Heroin Drought and its Implications for Drug Policy
No.58	Does prohibition deter cannabis use?
No.57	Firearms and Violent Crime in New South Wales
No.56	The Problem of Mobile Phone Theft
No.55	The Scope for Reducing Indigenous Imprisonment Rates
No.54	What Causes Crime?
No.53	New South Wales Drug Court Evaluation: Interim Report on Health and Well-Being of Participants
No.52	New South Wales Drug Court: Monitoring Report
No.51	Community Survey of Willingness to Receive Stolen Goods
No.50	New South Wales Drug Court Evaluation: Program and Participant Profiles
No.49	Drug Use Among Police Detainees
No.48	Performance Indicators for Drug Law Enforcement
No.47	Predicting Women's Responses to Violence: The 1996 Women's Safety Survey
No.46	Heroin harm minimisation: Do we really have to choose between law enforcement and treatment
No.45	The Effect of Arrest on Indigenous Employment Prospects
No.44	Public Perception of Neighbourhood Crime in New South Wales
No.43	Crime Against International Tourists
No.42	Predicting Violence Against Women: The 1996 Women's Safety Survey
No.41	Cannabis and Crime: Treatment Programs for Adolescent Cannabis Use
No.40	Are the Courts becoming more lenient? Recent trends in convictions & penalties in NSW Higher and Local Courts
No.39	Measuring Crime Dispersion
No.38	Mental Health and the Criminal Justice System
No.37	Crime Trends in New South Wales: The Crime Victim Survey Picture
No.36	Hung Juries and Majority Verdicts
No.35	Anabolic Steroid Abuse and Violence
No.34	Aborigines and Public Order Legislation in New South Wales
No.33	Child Neglect: Its Causes and its Role in Delinquency
No.32	Young People and Crime
No.31	'Home Invasions' and Robberies
No.30	Measuring Trial Court Performance: Indicators for Trial Case Processing
No.29	Methadone Maintenance Treatment as a Crime Control Measure
No.28	Crime Perception and Reality: Public Perceptions of the Risk of Criminal Victimisation in Australia
No.27	Risk Management in Assembling Juries

NSW Bureau of Crime Statistics and Research - Level 8, St James Centre, 111 Elizabeth Street, Sydney 2000
bcsr@agd.nsw.gov.au • www.lawlink.nsw.gov.au/bocsar • Ph: (02) 9231 9190 • Fax: (02) 9231 9187
ISSN 1030 - 1046 • ISBN 0 7313 2679 2