THE POLICE NATIONAL COMPUTER AND THE OFFENDERS INDEX: CAN THEY BE COMBINED FOR RESEARCH PURPOSES?

Brian Francis

Centre for Applied Statistics, Lancaster University

Paul Crosland

Institute of Criminology, Cambridge University

Contents

Executive Summary

Background

- The Offenders Index (OI) is a database intended to contain all court disposals relating to standard list offences since 1963 in England and Wales and is the standard research tool used for conviction studies.
- The Police National Computer (PNC) is an operational policing database for the UK (excluding Northern Ireland), which contains additional information, particularly on cautions, warnings and dates of offence. It is not suitable for research into criminal histories, as weeding of records takes place periodically. Police records from the period prior to its launch in 1995 are being computerised and integrated into the PNC database.
- The purpose of the present research is to investigate the feasibility of merging PNC records into extracts from the Offenders Index to maximise the information available for research and evaluation of patterns of offending.

Aim of research study

- To assess the quality of the PNC and the Offenders Index data sources
- For selected subgroups of cases, to match OI data with PNC data and to assess rates of matching
- To develop a strategy whereby, for any set of offenders, the two sets of information could be merged into a single data source

Data and Methods

- Five research datasets were examined in detail. For some of the datasets, names were collected from an external agency such as the probation or prison service and the names tracked on both the PNC and OI. For other datasets, offender names were collected from one source and traced on the other. The five datasets were : 1) Persistent Young Offenders, with individuals identified from PNC data; 2) Probation Pathfinder offenders from Devon, Hereford and Worcester, and Teesside; 3) Offenders used as the control group in research on the Prison Service's Home Detention Curfew scheme; 4) Offenders sentenced to any disposal in one week in 1997, as drawn from the annual OI 'Sentencing sample' and 5) Offenders from the Strategic Alliance Prospective Reconviction Study relating to four Probation Service areas.
- For each research study three files were supplied: a link file with data at the individual level and two files of court appearance records for the individuals in the link file, one obtained from the PNC and one from the OI.
- Offence codes and disposal codes on the PNC are different to those on the OI; and have also changed over time. Conversion routines currently available are not of sufficient quality to ensure accurate matching at the offence or disposal level.
- This research therefore focuses on matching records at the court appearance level.
- Purpose-written software was developed for linking and matching individual and court date level records, allowing manual intervention where necessary.
- The results of the matching process were used to develop an automatic matching algorithm.

Matching process

- For each study, the PNC and OI files were 'levelled' to ensure that information present on the PNC but not on the OI was removed before comparing records. This includes non-standard list offences, cautions, warnings reprimands and impending prosecutions, convictions before 1963 and convictions outside England and Wales.
- Using the purpose-written software the individuals in the link file were matched with the OI and PNC records on personal details such as name, date of birth and gender.
- If the personal details in the three files agreed (or partially agreed), then the level of matching at the court date level was tested. Two conviction records were accepted as belonging to the same person if the level of matching was high.
- User intervention involved comparing PNC and OI court and police identifiers together with summaries of the number of offences at each court date for each of ten offence groups. PNC and OI court dates were sometimes matched manually if they were close and other details agreed.
- Matches were, on occasion, 'partial', if the OI record appeared to contain composite information on more than one individual.

Results

- The matching process was carried out for 18,267 individuals over the five research datasets. Records for 16,814 (92%) were found in both the PNC and OI data.
- Of the 16,814 individuals with both PNC and OI data present, 16,405 (97.6%) were accepted as matches, 291 (1.7%) were rejected, and 118 (0.7%) were found to be partial matches.
- The records of 90.5% of individuals in the five link files were matched (or partially matched).
- The level of matching for women was lower than for men. At the individual level, the rate of matching for the all female Pathfinder Hereford and Worcester link file was only 67.1%. At the court date level, the difference in the rates of matching over all the studies, 71.4% for men compared to 63.8% for women, was statistically highly significant.
- There were 178,743 court dates corresponding to the individuals accepted as matches (or partial matches) over the five research studies. Some of these dates came from PNC records only (16.0%) and some from OI records only (13.1%).
- 69.1% of the court dates came from both the PNC and OI files and matched exactly; 1.7% were manually matched.
- For most of the studies, the match rate gradually improved over time, with the highest match rates found in the most recent time periods - either in the years 1990-1994 or 1995-1999. The overall match rate increased from 50.3% for the period before 1970 to 73.3% for the most recent period.
- An examination of whether the poorer matching in the earlier years is due to court dates missing from the OI or from PNC indicates that the two data sources contributed approximately equal numbers of unmatched records to the combined dataset. However, in recent years, it appears that (except in the Sentencing Sample) the PNC is contributing between 50% and 100% more court dates than the OI.

- When matching is examined according to the PNC's back record conversion indicator, the BRC indicators C, E, F and O are similar in the degree of matching (approximately 70%). The only low degree of matching in general is for BRC category N (21%). However, PYO produces a surprisingly high match rate for the N category (82.6%).
- Examination of the match rates by police authority after 1974 for the three larger studies (PYO, HDC, Sentencing sample) showed low rates of matching (averaging around 65%) for the Metropolitan Police area and for the City of London compared to rates of 80-90% for many Northern, Midlands and South Western forces.
- A comparison of the mean number of offences contributed by the two data sources for each matched court date indicates that before 1990 the PNC contributed a larger number of offences than the OI and that from 1990 onwards this pattern has been reversed. It must be noted that this finding is after 'levelling' the data to contain only standard list offences. The full PNC data, which includes non-standard list offences, will contain more offences by definition.
- The rates of matching for a number of common family names was often poorer than the rate of matching over all family names. One reason for this may be that the OI record for common family names is more likely than for other names to consist of composite individuals erroneously formed into a single record.

Automatic matching

- Using logistic regression on the results of the matching process on the Home Detention Curfew sample, a matching score was developed to link court level records from the PNC and OI. The matching score contained thirteen measures of potential discrepancy between the OI court level record and the PNC record. Greater discrepancy leads to a lower score – scores above zero for PNC-OI record pairs are accepted as matches.
- The matching score performed well on a validation sample from the same dataset, with only 0.35% of records mismatched when compared to the 'true' results obtained from the earlier process.
- The matching score also performed well on a validation sample from a different dataset (OI sentencing sample), with only 0.40% of records mismatched when compared to the 'true' results obtained from the earlier process.
- Based on these results an automatic matching algorithm has been proposed which will have minimal user intervention.

i) Report Recommendations

- Work should be undertaken to incorporate PNC records into the Offenders Index as set out in this report. This will increase information on known offending for research and evaluation purposes as suggested in the Review of Statistics on Efficacy of Sentencing. The gains offered by the addition of PNC data can only be fully realised by including all types of disposal and all types of offence (rather than just Standard List offences). (*Section 1*)

(High priority)

- That statistics are produced showing variation by Police Authority in the degree of matching achieved with court records on the Offenders Index. Statistics could also be produced from the enhanced Offenders Index showing the time gap between offence and sentence, and in pseudo-reconviction rates by type of disposal. (*Section 12*)

(Medium priority).

Augmentation of existing OI records

- That existing OI records be augmented by matching PNC information at the court date level. This information would include a summary of PNC offences and disposals as follows:
  - total number of offences
  - number of offences in each of the 10 categories of offence
  - number of offences committed on Bail.
  - earliest and latest offence start dates and offence end dates known
  - a PNC disposal date (which may differ from the OI disposal date)
  - PNCID (Section)
  - Co-offender PNCIDs, CROs and Dates of Birth
  - Local Police Force and Police Station Post Code values

(*Section 11.2*)

(High priority)

Adding new disposal records to offenders traced on the OI

- That additional court dates and other disposals from the PNC be added to the OI, including both cautions/warnings/reprimands and impending prosecutions and also court disposals not known to the OI. *(Section 11.1)*

(High priority)

Creating new records for individuals on the OI from PNC information

- That new records are created from PNC information for individuals not traced on the OI. (*Section 11.1)*

(High priority)

De-merging incorrect OI records

- That OI records incorrectly merged be systematically 'de-merged' once the PNC dataset has been added to the OI. (*Section 10.2*)

(Low priority)

Coding

- That RDS develop comprehensive, reliable, regularly updated and transparent coding lists and syntax to enable offences from the PNC and the Offenders Index both to be identifiable and translatable in the following ways:
  - Statutory legislation establishing/ending the offence
  - ACPO code (to as many levels as exist at the time of the Offence)
  - CCCJS code
  - Offenders Index Code (with dates of application)
  - Categorisation into eg 10 categories of offence (with dates of application)
  - Standard list/non-standard designation (with dates of application)
  - Development of a new code created for matching PNC and OI data at offence level which does not require reference to dates for interpretation.

  (*Section 5.1*)

(Medium priority)

- That RDS undertakes the same task in relation to the PNC and OI disposal categories. (*Section 5.2*)

(Medium Priority)

Offence level matching

- That, following the above overhaul of coding, research is undertaken to review the robustness of the offence level match, using similar techniques to those developed in this report

(*Section 8 3*)

(Medium Priority)

- That consideration be given to a request for the PNC to make offence start date a mandatory data field. (*Section 11.2*)

(Medium Priority)

Conviction research and evaluation studies

That RDS develops best practice guidelines for evaluation studies and other conviction research based on the enhanced dataset, describing the use of the new information. (*Section 13*)

(Medium priority)

Supply of data

- That RDS improves its service to researchers in ensuring that every data file (OI, PNC or hybridised) is provided with an electronic version of the data structure and all coding used

(with the latest date of change of the data structure and any of the coding recorded). Data positions of researchers additional fields should be provided (*Section 13.6*)

(Low priority)

- That additional research be carried out to further develop, test and evaluate the proposed automatic matching of records at the court date level and the automatic detection of OI criminal histories which need demerging. (*Section 14*)

(High priority)

- That RDS implement the results of the research above into purpose written software capable of matching large numbers of criminal records together efficiently. *(Section 14*)

(High priority)

1. Introduction

The primary purpose of this research was to investigate the potential for adding value to the current source of offending data for research (the Offenders Index or OI) by augmenting it with data from the operational Police National Computer (PNC). Extracts of information from the PNC files can enhance research studies by providing dates of offences, information on cautioning and warnings and other relevant information. PNC data can also sometimes enable research studies to be undertaken more promptly in advance of Offenders Index data being available, for the Offenders Index is updated quarterly, six months in arrears.

1.1 Background to the research

The 'Review on Efficacy of Sentencing' (Allnutt, 2001) is the starting point for this research project. In his report, Allnutt suggested the creation of a 'Complete criminal record', with information being provided from the Police, the courts, the Prison Service and the Probation service to make a complete criminal history, containing dates of offending, sentencing, periods of custody, dates of release and other pertinent information. This research project is one small step towards that ambition and investigates the possibility of bringing together PNC data and OI data. Incorporating PNC data into the Offenders Index will have the following immediate advantages:

a) it will encourage research based on the PNC-recorded dates of offence rather than on sentencing or court dates. This will remove the problem of pseudo-reconvictions where an offence is committed before a target date (a date of sentence) but not convicted until after the target date.
b) It will introduce non-court disposals into the OI. This will include cautions, reprimands, warnings and final warnings. This could be seen as overdue, as there is increasing effort –particularly in youth justice- in developing complex interventions that do not involve the courts e.g. Police force restorative conferencing at Final Warning stage.
c) It will introduce non-standard list offences into criminal histories. One use of this would be into research on serious sex offending, where minor offences such as indecent exposure without intent to assault is viewed by some as a precursive risk factor to becoming a serious sexual offender; investigation of this hypothesis would become possible.
d) It will be much easier to undertake further research to review the completeness of conviction records in each data system.

It has been suggested by Allnutt (2001) that the PNC might in time become the main repository for research and statistical information, but this is presently an unrealistic prospect, and we have not pursued this suggestion in this report. The PNC currently lacks an interrogation tool i.e. a means of extracting records based on common criteria rather than individually retrieved PNC identifiers, and the protocols for access to information would have to be thoroughly explored. A protocol has been set up between the Research, Development and Statistics Directorate of the Home Office (RDS) and the Police Information Technology Organisation (PITO) to extract small files relating to known PNC identifiers. This is one building-block towards the complete record access aspired to in the Criminal Justice system.

Comparing datasets derived from different parts of the Criminal Justice system helps to highlight some of the weaknesses of each system and begins to address whether the production of a hybridised dataset is achievable and what advantages such a dataset would have. The consequences of errors in the matching of one individual with another individual's criminal record are clearly different according to the use of the resulting dataset. We emphasise that the result of the matching process should only be used for research studies, where an occasional mismatched record will have little effect on the broad population patterns. The merged dataset should not be used as an operational criminal justice system where inaccurate merging is highly problematic. Offenders will sometimes try to thwart the criminal justice system by giving different personal details (names, dates of birth etc) on different occasions, and this adds to the inevitable data-entry mistakes that any data system suffers.

Every correct record known to the Offenders Index should be recorded on the Police National Computer–at least for court sentences subsequent to the 1995 reincarnation of the PNC (Phoenix) when each Police Force was given this specific responsibility. Variation by Police force and date in the extent to which PNC and Offenders Index records match is not surprising in the light of other reports written exclusively about the PNC (Russell, 1998 and Povey, 2000).

Finally, this research project will develop methodology that not only seeks to quantify the apparent gaps in each system but will aid subsequent research using merged data which the Home Office undertakes or facilitates into offenders.

1.2 Definitional issues

The Offenders Index is the data source that has historically been most used, covering standard list offences (with some exceptions) reaching 'completion of processing' in any court in England and Wales. The word '**conviction**' has been used too loosely to define the court records which are added to the Offenders Index, quarterly and six months in arrears. Whilst the date of conviction is often the date recorded within court records, frequently the date that reaches the Offenders Index is a date subsequent to the conviction date. Where sentencing occurs in the same court on a date other than the date of conviction, the protocol is that this is the date that the court records pass to the Offenders Index. Gaps between the date of conviction and the date of sentence can be of great consequence both for research findings and for the matching of records holding different date information.

Using the Offenders Index and the PNC, the dates available for comparison are what we shall refer to as the **court dates**; rather than 'dates of conviction'. We retain the phrase '**time to reconviction**' as shorthand for measuring gaps between court and non-court disposals and also **gaps** between court disposals and reoffending. Being able to match 'court dates' is a pre-requisite to the matching of offences for a given court date.

1.3 The research project

This research has proceeded by comparing two sources of data; the Offenders Index and data files obtained from the Police National Computer. The Offenders Index (OI) is a database intended to contain all court disposals relating to standard list offences since 1963 and is the standard research tool used for conviction studies, answering certain Parliamentary Questions and constructing yearly information in sections of Criminal Statistics. The Police National

Computer (PNC) is an operational policing database, not designed to meet research requests and not retaining all records in perpetuity.

This report identifies differences in the information known to each system. Although PNC data contains more information than OI data, it is not designed for research purposes and weeding of records takes place periodically. We therefore advocate that hybrid research datasets be created by bringing PNC records systematically into the OI. This would improve research quality into the 'effectiveness' of various Criminal Justice System interventions by providing additional information, particularly on cautions, warnings and dates of offence.

It is important to emphasise that the augmentation of offending records described here is to be used for research studies; the aim is not to augment the operational PNC system with additional information. The occasional mismatch of records is acceptable in a research context but would not be acceptable for operational use. Thus, the direction is important – extracts from PNC can be brought into the OI, and not the other way around.

The detailed aims of this research study were as follows:

1. To assess the quality of the PNC and the Offenders Index data sources, using rigorous measures of assessment on a variety of criminal career information, including but not limited to evaluation studies.
2. For selected subgroups of cases, to match OI data with PNC data and to assess rates of matching.
3. To develop a strategy whereby, for any set of offenders, the two sets of information could be merged into a single data source.

The first aim seeks to find and quantify anomalies between given datasets. The datasets supplied by the Home Office to the researchers have unique personal identifiers within them e.g. the PNCID, though for any given individual, the possibility of that individual having a split criminal record (i.e. two or more IDs) has not been accommodated. Clearly for the Home Office to supply an entire copy of the many millions of records on the Offenders' Index and the Police National Computer would have been beyond the scope of the present project. The research is therefore not reporting back on the definitive existence of records being held on either database, but only on whether the retrieved records (based on one ID per person) match where they should match and beyond that, whether the records can provide useful supplementary data to enhance research on offenders processed within the criminal justice system.

The majority of the effort in this study has been spent in developing methodology. To aid this process, five separate research studies supplied data to this project which enabled us to compare and match PNC and OI information. To undertake this comparison, a number of tasks needed to be carried out:

1. the levelling of PNC and OI files to ensure that information present on the PNC but not on the OI is removed before comparing the degree of matching. These included non-standard list offences, cautions, warnings reprimands and impending prosecutions, convictions before 1963 and convictions outside England and Wales.

2. the development of purpose-written research software for matching individuals and court date related details, allowing manual intervention where necessary.

3. The use of the results of stage 2 to develop an automatic matching algorithm.

The report proceeds as follows. Following a discussion of the research design (Section 2) a description of the PNC and the OI data collection processes is supplied (Sections 3 and 4), which is followed by a discussion of coding issues (Section 5). The five research datasets are then described (Section 6), and this is followed by methodological details of the levelling (Section 7) and matching (Section 8) process. Following a number of case studies indicating the matching system in operation (Section 9), the results of the matching are then given in detail. This is followed by a proposed structure for the incorporation of PNC records into the OI and some comments from the researchers about the needs of research into criminal justice system disposals and the offences to which they relate (Sections 11-13). The report concludes with a description of an automatic algorithm to match OI to PNC records (Section 14)

2. The research design

2.1 Models of how records are retrieved

The five datasets being examined in this research consists of pairs of PNC and OI files relating to the following studies:

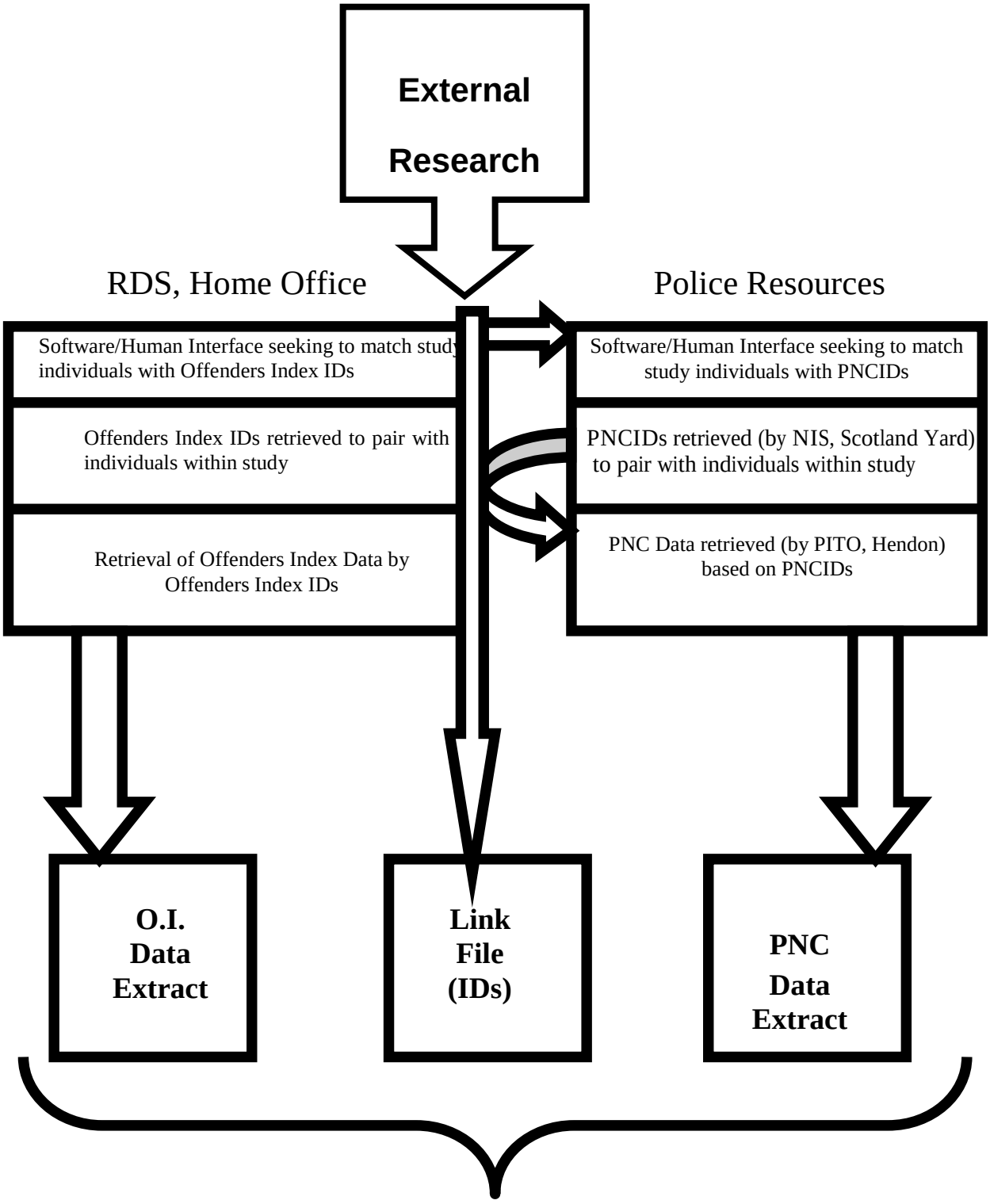
1) Probation Pathfinder Study (sub-divided into 3 geographical datasets)
2) Home Detention Curfew Study
3) Strategic Alliance Reconviction Study
4) Persistent Young Offenders Study
5) 1997 Sentencing Sample

Figures 2.1 and 2.2 show how the records for the research datasets have been retrieved. There have been two models of data retrieval used. The first model (Figure 2.1) has taken an *external data source* as the starting point, with names of offenders being passed from that source to both the OI and the PNC for tracing. The external sources used here were the Prison Service (Home Detention Curfew) and the Probation Service (Probation Pathfinder and Strategic Alliance studies), and this model is typical of the way in which the OI is currently used for evaluation.

The second model involves selecting cases which conform to some set criteria on one of the datasets and tracing the names found on the other(Figure 2.2). Thus, the 1997 sentencing sample selects from the OI all those individuals who have been sentenced on one of a number of days in 1997. This model is typical of more general research into offending, where the aim might be to examine criminal history patterns of murderers, sexual offenders or fraudsters.

It has not been possible however to explore other measures of comparing datasets (Figure 2.3). For example, a more valid comparison of the 1997 OI sentencing sample would be to query the PNC for the criminal records of those sentenced on the same days in 1997. This could not be carried out, as such a querying facility within the PNC has not been part of policing requirements, and so does not exist. The closest which has been achieved is the coding framework that has been developed (at some expense) for the Lord Chancellor's Department so that numbers of persistent young offenders can be monitored monthly. Besides being able to extract persistent young offenders from the PNC, the only method by which the Home Office RDS Directorate can retrieve records is by supplying known PNCIDs (identifiers for individuals within the PNC). Obtaining PNC data in this way has made the methodology of comparing data more convoluted than if parallel retrieval of cases meeting certain criteria had been obtained from both data sources.

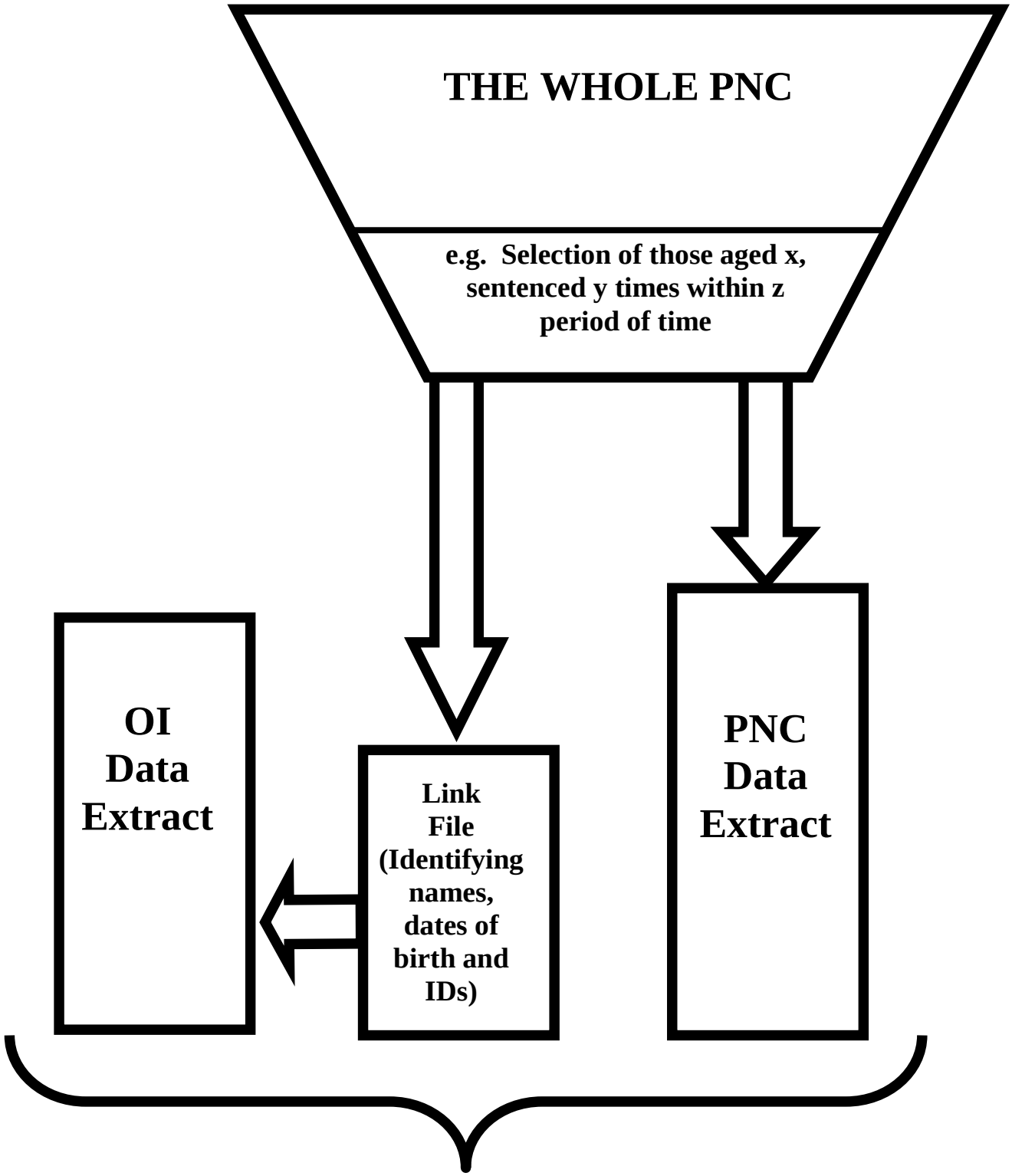
Figure 2.1 -Data source external to PNC and OI



Model 1 data extraction applies to the Following Robustness Research Datasets:

- Pathfinder -Teesside, Devon, Hereford & Worcester (originating from researcher records)
- Home Detention Curfew dataset (Prison records)
- Strategic Alliance Reconviction Study (Probation records).

Figure 2.2 Retrieval from one data source based on internal criteria



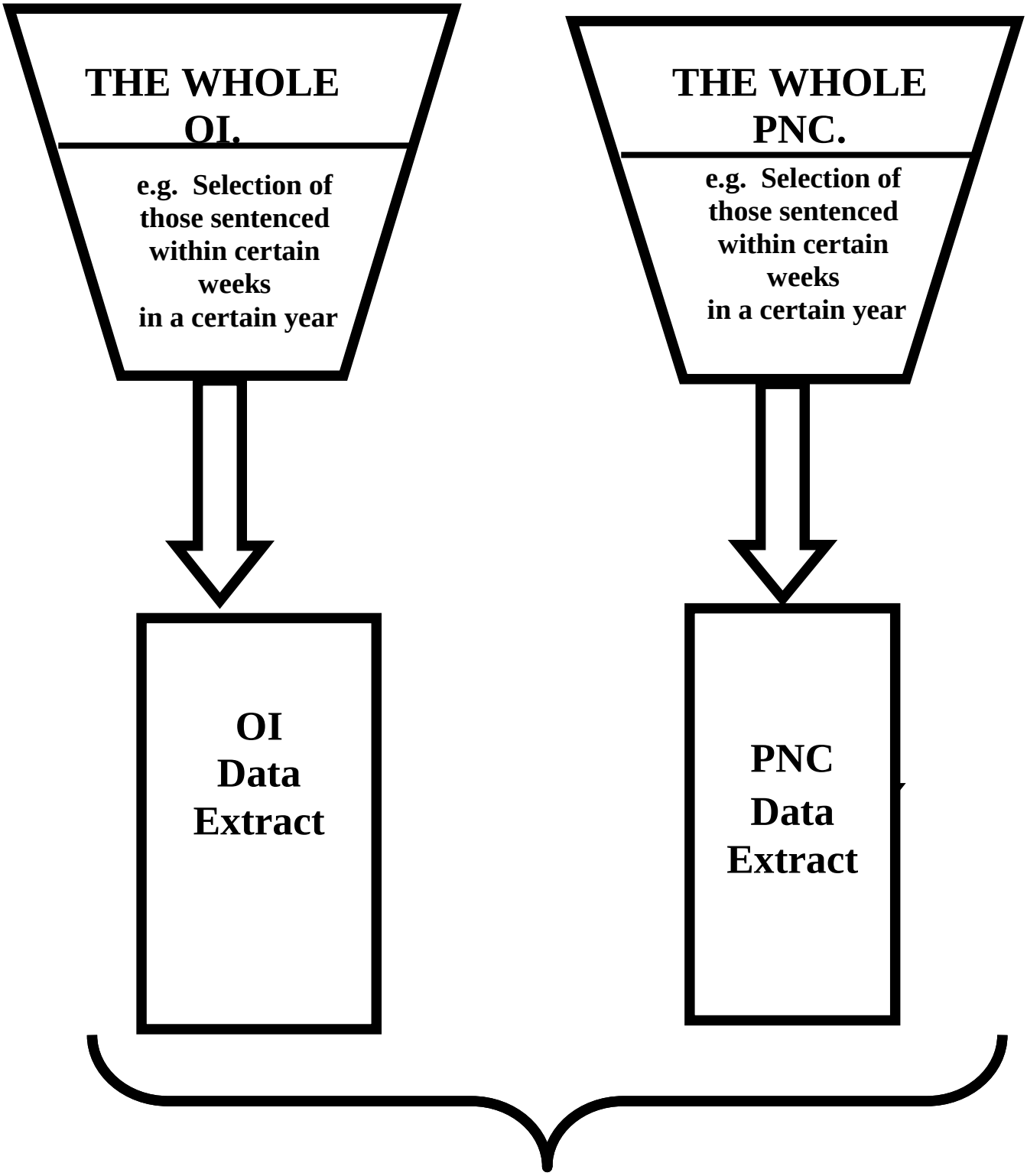
Model 2 data extraction applies to the Following Robustness Research Datasets:
Source data –PNC

- Persistent Young Offenders identified within PNC.

Source data –Offenders Index

- Sentencing Sample (from 1997).

Figure 2.3 Idealised but unobtainable parallel retrieval of records based on internal criteria



Model 3 data extraction applies to none of the Robustness Research Datasets.

2.2 Offences, court cases and the supply of data to the PNC and the Offenders Index

Figures 2.4 and 2.5 shows in schematic form the process by which data is supplied to both the PNC and the Offenders Index from a court case. The difference can easily be seen – the Offenders Index is only updated from the court databases six monthly in arrears, whereas the PNC data can be updated immediately by processing information supplied by the courts. These figures also highlight other issues, such as the weeding of PNC data of less important offending information, and back record conversion, which is the process of computerising microfiched records from the period prior to the 1995 launch of the PNC as the data system on which all Police Forces were given the responsibility of entering data. (Back record conversion is particularly an issue for records that commenced prior to 1981.)

Figure 2.4 shows that the route by which records end up on the Offenders Index involves (at the time of writing) approximately half the Police Forces forwarding the Magistrates Court details to the Court Proceedings Database at the Home Office. The Magistrates Courts in the remaining 20 forces have a direct supply of data to the Court Proceedings Database (Figure 2.5). More Police Force Areas will lose the role of supplying the Court Proceedings database as the Magistrates Court computer system is extended, though all Police Forces will still have to update the PNC with the cases from their local Magistrates and Crown Courts. (All Crown Courts supply data directly to the Court Proceedings Database).

Figure 2.4 How records are created –20 Police Forces (as Appendix 2).

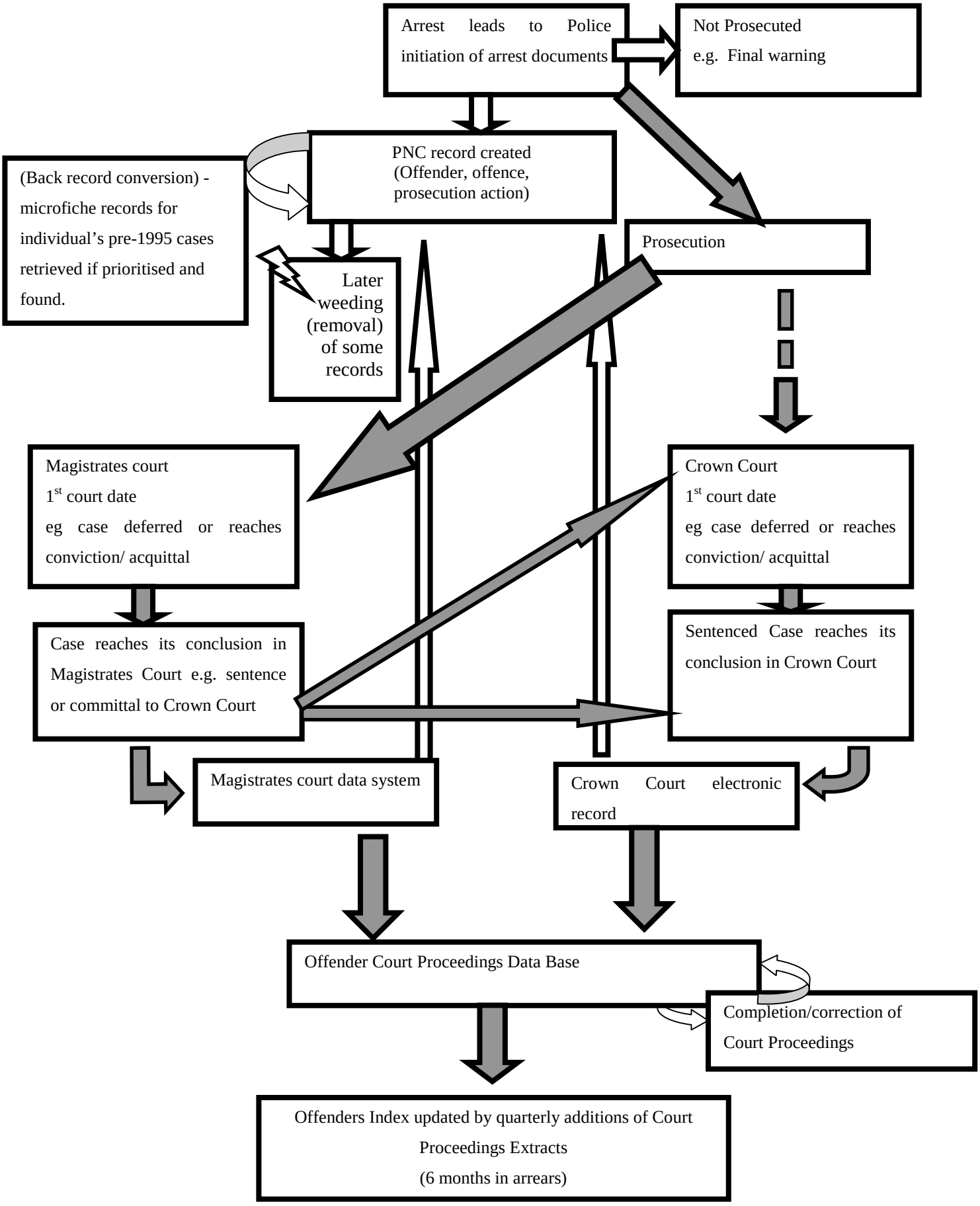
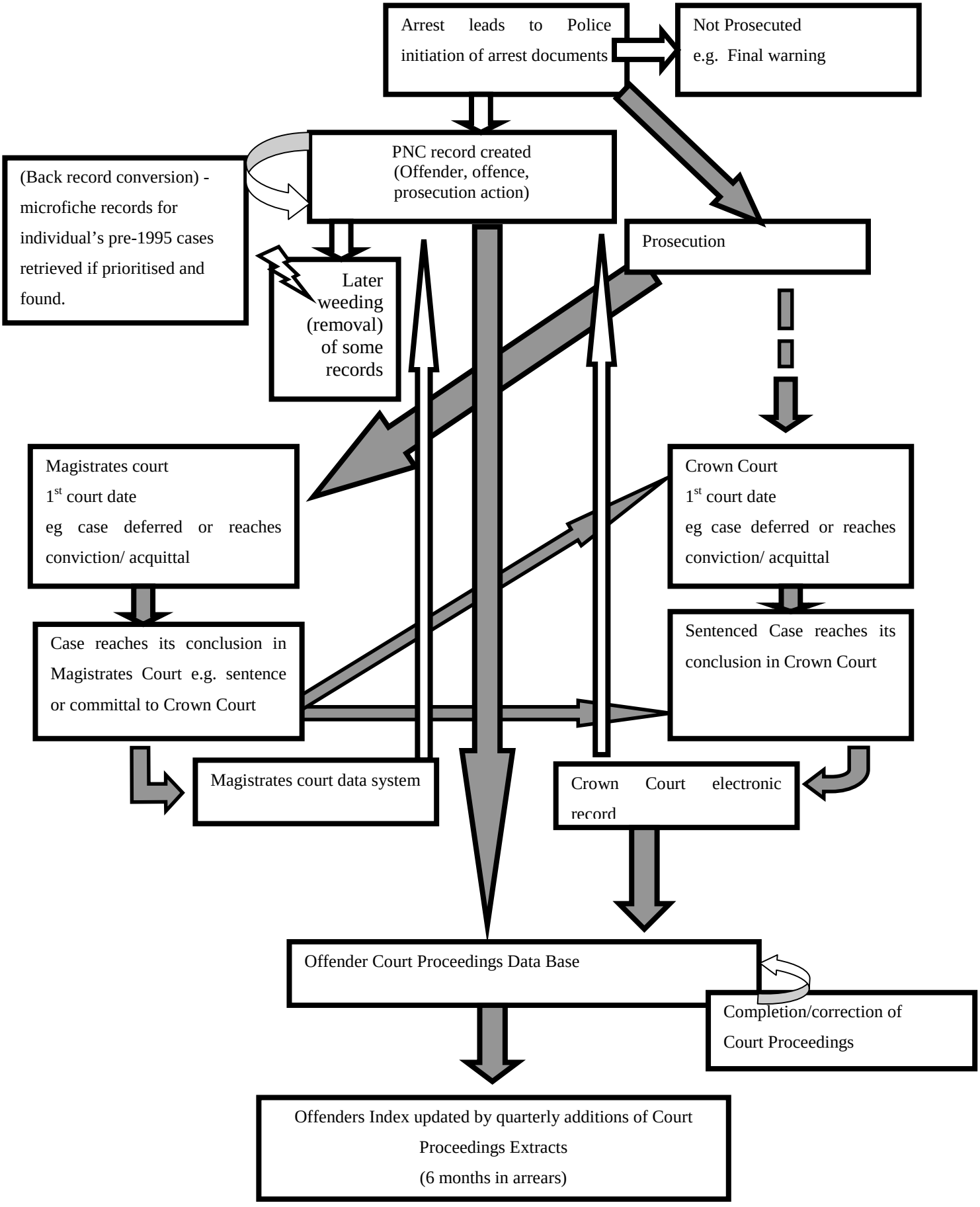


Figure 2.5 How records are created –23 Police Forces (as Appendix 2).



3. The PNC data

The PNC is an operational police database containing criminal histories of all offenders in England, Wales and Scotland- it also contains information from other police forces such as the British Transport Police. Criminal records prior to 1995 need to be back record converted, and although this process is well under way, it is not complete. However, when records are converted, then the criminal history is complete, and not limited, as the Offenders Index is, to convictions after 1962. It contains much potentially useful information for research purposes. The emphasis in this report is on the availability of dates of cautions, reprimands and warnings, and dates of offence. However, information on postcodes of offences is also potentially available.

The PNC, however, is not, and cannot be, the optimum repository of criminal histories for research purposes. It has not been designed for this purpose and it is also committed to a (recently tightened) policy of 'weeding' records. This deletion of records removes records no longer of interest to the police, meets Data Protection needs and respects the interests of rehabilitation of offenders. New weeding rules will remove the records of those offenders who have not been convicted within the last 10 years (previously 20 years) nor cautioned within the last five years "and who satisfy a number of other conditions concerning age and previous offences, disposals and mental health status. Records are also deleted one year after the death of an offender. Long-term follow-up studies will be affected in particular." (Howard and Kershaw, 2000).

However, bearing this in mind, the PNC contains much useful information, particularly on recent criminal histories, which can be used to give added value to research studies based on the OI.

3.1 PNC file levels

The PNC data was supplied using a standard file transfer protocol established between the Home Office's RDS and the Police Information Technology Organisation (PITO). Much of the detail of the PNC has been suppressed in the resulting data file as it is unnecessary for research purposes. The data structure supplied has six record types –as opposed to the three types of the Offenders Index.

Type 1 –OFFENDER details
Type 2 -PROCEEDINGS details
Type 3 -OFFENCE details
Type 4 –COURT DISPOSAL details
Type 5 –SUBSEQUENT APPEARANCE details
Type 6 –CO-OFFENDER details

Unlike a file obtained from the Offenders Index, a PNC data file usually contains many proceedings records without court disposals; that is, many proceedings (type 2 records) will have no type 4 records. Whereas type 4 records are only for court outcomes, type 2 records also capture details of impending prosecutions, cautions, reprimands, and warnings where there are no court outcomes. The dataset is hierarchical up to Type 4. Thus for each Offender record (Type 1) there can be many proceedings records (Type 2); for each proceedings there can be one or more offence records, and for each offence record there can be none, one or many court disposal records. Subsequent appearance details (type 5) occur less frequently than type 4 records. Co-offender details are supplied as another type (6) though

this is not a hierarchical data arrangement in that co-offenders being recorded is not dependent on the existence of type 4 and type 5 records.

The rather complex structure of the dataset creates problems. The most popular research software within the criminological research community (SPSS) is unable to read in the PNC data file structure beyond type 3. This is to say that if –for example- the SPSS programme is set to read in four levels of PNC data the software will only build records for those individuals who have a level 4 record (disposal following conviction). Thus, no records for those cautioned or warned would be built. For this research, a simple piece of software (using an `awk` DOS-based script)[1] was written to post-process the datasets. This script, in the absence of a level 4 record, automatically writes a blank level 4 record into the raw data file to enable the SPSS four level NESTED import not to drop those records with only three levels. For future studies where PNC data is passed directly to researchers, we recommend that blank records at levels 4, 5 and 6 be entered where none exist – this will remove the need for an `awk` script or other post-processing of the raw file.

3.2 Stage 1: Obtaining PNCIDs from personal information

The first stage in obtaining a PNC dataset is to obtain PNC identification numbers (PNCIDs) from personal information. Details (usually surname, initials and date of birth, plus sometimes a date of sentence known to the researchers) are provided to National Identification Service at Scotland Yard (NIS) who retrieve the PNCIDs from the dataset using standard PNC search software. The results presented in Section 10 show that on average, 6.2% of offenders are not traced at this stage. The results vary from research study to research study – the Home Detention Curfew Study which samples prisoners had least difficulty, with only 0.1% of names untraced.

This research has had to take the retrieved PNCIDs as the starting point. It appears that the response 'No trace' is given when there may be possible imperfect matches in existence. The authors have used both previous research experience and a visit to a police station to observe the retrieval system in operation and have made suggestions for improvement.

a) Providing the following additional details on the offender would improve the accurate retrieval of PNCIDs (though none of these items should be compulsory for a match to be ascertained):

- sex of offender;
- full first names;
- CRO number where known;
- a geographical area where the individual is known to have been living recently/ sentenced/born.

b) Where no 'perfect match' can be ascertained, allowing for the possibility of returning (and RDS handling) more than one PNC record for the offender, with levels of confidence for each match. (With common family names, there may also be more than one perfect match).

[1] `awk` is a unix-based fast record processing utility. `gawk` is a free software foundation version of `awk`. See `http://www.gnu.org/software/gawk/gawk.html`

For example, in one of the research datasets (the Strategic Alliance Reconviction study) there were eleven out of 1400 cases that were returned as 'No Trace'. One of these records was instantly matched by the PNC software when surname, initials and date of birth were entered. (There was but one letter difference in the middle of the surname.)

The current operational system which only supplies one PNCID and thus only one record to the researchers is likely to lead to many fewer offences for an individual being put into research than are actually held on records by the criminal justice data base. To quantify the level of split records on the PNC would require a fresh research project and it is only due to other research undertaken by the authors that clear case examples of split records have been stumbled across. We recommend that the beneifts of NIS returning more than one PNCIDs be examined.

Another problem relates to split records on the PNC. Whilst the PNC has software enabling an offender for whom fingerprints have been obtained to be joined to another with the same fingerprints regardless of name or date of birth used, there are a number of circumstances where offenders have split records on the PNC. Many of these appear to be records which have not been reconciled by the police. PITO can produce printouts of duplicate records as known to them, which may help RDS cope with the complexity of not being able to presume that all individuals have just one PNCID over time.

Caution on the part of the Police in joining split records together is very understandable. An operational police system needs to take great care in bringing together such records, whereas researchers may weigh the cost and benefits of some incorrectly merged records versus many split records quite differently. More research could be undertaken in this area to get a clearer idea as to the size of the problem.

3.3 Stage 2. Obtaining PITO data from PNCIDs

Once PNCIDs have been obtained they are passed to the Police Information Technology Organisation (PITO) for retrieval of records in a standardised data format. The 'basic unit' of the data retrieval could be said to be the offender/offence. For each offender/offence the number of data fields that will be retrieved theoretically varies between the 7 mandatory fields for records type 1 and 3 and the 43 fields provided by the data protocol arranged with PITO. The PNC itself contains a total of 233 data fields relating to offender/offence. It has been possible to obtain a volume of documentation relating to the 233 fields held on the PNC (Police Information Technology Organisation, 2000). This four hundred page document is itself incomplete as it refers to ACPO data standards and codings contained elsewhere. The 43 field names supplied by RDS, however, do not always match those in the Data Definitions documentation and for some fields, the information has been combined from two or more existing fields.

Those fields that are important enough to facilitate mainstream (or specialist) criminological research should be subject to periodic review and negotiation. Examples of data fields which would help in improved retrieval and matching of PNC records include alias names, Arrest Fingerprint Status, and Arrest Date of Birth, the latter described as :

> *"The DATE OF BIRTH given by the subject at time of arrest/charging. It may be the same as the file Date of Birth or any Alias Date of Birth already recorded or may be a new Date of Birth. NOTE : Where the given Date of Birth is not already recorded the file Date of Birth (for a new subject) or a new Alias Date of Birth (for a recidivist) will be created from this data item" (*ibid*).*

The quality of recording can valuably be reviewed as well; for example, there appears to have been an increase in ethnicity recording following the Lawrence Report. Monitoring the progress of the use of the PNC fields can be vital to an understanding of their meaning.

The fields which are mandatory for completion on the PNC and which are provided in the RDS-PITO protocol are provided as Appendix 1 to this report. Some fields which may be useful to researchers currently have a non-mandatory status. For example, for research involving court sentencing the 'disposal qualifier' field is non-mandatory. This field records for example whether a sentence was concurrent or consecutive. Estimation of the effective length of a sentence from the PNC where there has been no use of the 'disposal qualifier' field is thus impossible.

3.4 PNC data quality and timeliness

The completeness of the PNC data is a crucial element of this study, and one that has been addressed by previous reports. Examples of incomplete PNC records have been cited in the Police Research Group's 1998 report:

> *"We reported no previous history for the offender. We subsequently discovered he had six previous convictions sitting in our backlog of records. But the offender had already been bailed."*
> *"Our backlog was so bad, judges were complaining the force's records were not up to date. In one case the Superintendent was called to account in court. The offender had served three years for rape and this had not yet been recorded") (Russell,1998,p33)*

Her Majesty's Inspector of Police reported in 2000 that he

> *"considers the level and nature of errors, omissions and discrepancies found to be totally unacceptable especially given that many of these same observations were made in the 1998 PRG Report. They reflect an unprofessional approach to data quality by forces" (Povey, 2000, p142).*

In 2001 more pressure has been applied to try to improve the situation:

> *"The data Protection Commissioner, has told police and the Home Office that she is prepared to use her powers to order forces publicly to clear up huge backlogs of convictions and court results on the Police National Computer system….Many hundreds of thousands of records needed updating with results of court cases stretching back months. In one force the average time to process cases was 413 days…Emergency plans [to stave off Data Protection enforcement orders] are being supervised by Home Office inspectors of constabulary who are visiting forces every two weeks." (Tendler, 2001).*

The completeness of PNC records is subject to fluctuation according to the priorities of the police recording procedures, just as the Offenders Index is subject to the accuracy of court record updating. The OI is updated quarterly, six months in arrears from court data. Whilst Court Proceedings and Cautions reports are produced by the Home Office Data Collection Group, the relationship of OI data to PNC data has not –until now- been reviewed in a Home Office

research project, though Crosland (1999) and Friendship (2000) have raised cause for concern and have suggested that further research is needed.

From the HMIC report it is also possible to glean something of the history of the PNC data gaps and the attempts to ensure that the PNC data becomes more complete in the future:

> "*Overall Her Majesty's Inspector considers the Record Type and nature of errors omissions and discrepancies found to be totally unacceptable especially given that many of these same observations were made in the 1998 PRG Report. They reflect an unprofessional approach to data quality by forces." (Povey, 2000, p142)*

The ACPO Compliance Strategy for PNC (the National Phoenix Performance Indicators) state that all court case results should be entered within 72 hours of coming into police possession. From the PNC Performance Statistics published by PITO it is possible to monitor some elements of the timeliness of PNC record entry; supplying this supporting information on the changing timeliness of PNC records is vital to research which tries to be up to date.

The following figures were made available at the start of this project (October 2000 figures)

- The time taken for English Police forces to record 90% of arrest/summons reports varied by police force from 1 to 125 days, with a mean of 36 days. (A year previously the average was 45 days). (In Wales the average of 63 days was unchanged).
- Days (from Court Date) for English Police forces to record 50% of court results varied by police force from 7 to 101 days, with a mean of 29 days. (This is an improvement of 4 days over the October 1999 average figure for English Police Forces. In Wales the average declined from 24 days to 37 days).
- Days (from Court Date) for English Police forces to record 90% of court results varied by police force from 21 to 612 days, with a mean of 206. (This shows a decline of 10 days over the October 1999 average figure for English Police Forces. In Wales the average declined from 162 days to 183 days)

It is however too easy to read these court date figures too negatively. The impending prosecution 'court date' initially recorded is very often not the date on which the case reaches sentence so to interpret these gaps as the gaps between court outcome and police recording of court outcome is often incorrect.

Other research methodologies could be applied to ascertain delays in recording sentences. For example, obtaining two data files from the PNC at different points in time and comparing them is one possible methodology which would yield data on changes over time.

3.5. Back record conversion.

Back record conversion is the process of computerising microfiched records from the period prior to the 1995 launch of the PNC, and is particularly an issue for records that commenced prior to 1981. The PITO web pages show pride in the progress of back record conversion of data:

> *"The BRC project team achieved a significant milestone in 1999 when they converted their 500,000th criminal record from microfiche onto the Phoenix database"*.

It seems there is no short or medium term possibility of getting the PNC to have the historical strength of the Offenders Index –with its systematic collection of data since 1963. The task of converting all incoming offenders old enough to have records on microfiche (anyone born before 1985) is just too large to be prioritised amongst the apparent operational concerns of the Police force. The Police priority in their use of the PNC is to record arrests quickly; catching up on court records (or microfiche) lags behind, as evidenced by the PITO statistics.

3.6 Advantages and disadvantages of the PNC

Advantages	**Disadvantages**
Includes Scotland, England, Wales, British Transport Police.	Cannot search the database directly
Complete history for older offenders if back records are converted	Offence codes and disposal codes differ from standard RDS codes (see Section 5)
Information usually available more promptly, but delays in data entry.	All offences recorded
Criminal histories built up by fingerprint verification	Criminal histories might be split.
Cautions, warnings and impending prosecutions available	Criminal histories weeded – less important offences removed
Dates of offence available	disposal information of poor quality
	Criminal histories deleted on death

4. Offenders Index data

The Offenders Index provides a history of criminal convictions from 1963 using data collected from England and Wales Crown and Magistrates courts. It contains information on sentencing dates, age, offences successfully prosecuted and sentencing (disposal). It does not deal with all offences but instead records all standard list offences. As it is a court – based system, there is no information on dates of offending; and information on dates of custody and release are also not present. There is also no information on unsuccessful prosecutions and cautions. Information on appeals is not added, and thus some convictions on the index may have been overturned by the court at a later occasion. The Offenders Index is updated quarterly, six months in arrears. In short, conviction research using the Offenders Index automatically has a minimum of six to nine month research time lag.

A major advantage of the Offenders Index has been the ability to query the database directly. In the past, this has been used to carry out studies of all those convicted for a certain offence in a certain period of time (see e.g. Soothill et al, 2000, Rose, 2000), studies of sentencing behaviour in a certain year (Home Office, 2000) and extractions of birth cohorts of offenders (Prime et al , 2001). This latter dataset has been anonymised for academic use and is available at the UK data archive.

4.1 The OI database

There are three file levels to the dataset. The dataset has a hierarchical nature, with offences (level 3) nested within court appearances (Level 2) which in turn is nested within individuals (Level 1). The Offenders Index Users Guide provides detailed information on the format of the records and the information provided.

Level 1 –OFFENDER details
Level 2 -PROCEEDINGS details
Level 3 -OFFENCE details

Disposal information is provided on the offence details; for each offence there is a maximum of four disposal slots available. It is rare that all four disposal slots are used.
The dataset is well-documented by researchers, and documentation is available for reading supplied datasets into SPSS and other packages.

4.2 The RDS retrieval of OI data

External searches begin with the user passing surname, initials, date of birth and gender to the Offenders Index. Researchers may also pass additional fields to facilitate matching; it is not known whether this was done in these studies. The file of information is then passed to the matching software SSA-NAMES.

The Offenders Index user guide states that

> *"The new software automatically matches and assigns a computer-generated (and unique) OI number to a study offender with a record on the database if the surname, initials, date of birth and gender all correspond precisely. A score is given to all*

possible matches and staff are required to consider those above or below thresholds which are set according to the nature of the study. Staff are able to deal with possible duplicate offenders as they arise and are required to 'merge' as they work through the study. Merging takes place when it is apparent that an offender has been entered on the database twice or more and it is necessary to reconcile the criminal histories by merging to produce a complete record.

This is rather a confusing statement, and does not fully describe how the matching works in practice. A better description is that the scoring scheme is used to score near misses and is used when a precise match can NOT be achieved.

4.3 Known issues with the Offenders Index

The first is what is commonly known as pseudo-reconvictions, where, following a target conviction, the interest is in new offending following that conviction. New convictions can appear on the database following the target conviction in time, but which relate to offences committed before the target conviction. Thus, the time to the next conviction may not be a reliable proxy for the time to the next offence. However, this is not an issue in this study, as the date of conviction is not wrong.

Another issue relates to deferred sentences. Where a sentence is deferred, two records often appear in the database, one indicating that the sentence is to be deferred, and one for the actual sentence awarded at a later court appearance. Simplistic use of the Index would treat the second entry as a reconviction. Again, such records need to be identified. It would also be of interest to see what happens to such records on the PNC database.

Changes in the coding frame and the coding scheme over the years has also further complicated the tasks involved in processing the data. This is very much an issue for this study. There are three separate types of changes:

1. Incremental and gradual changes in the coding scheme over time. This has happened to the offence codes. As indictable offences are limited to codes between 1 and 99, the codes have been reused over the years. So, for example, code 27 represented sacrilege before1971, and soliciting by a man after 1979. Code 58 represented forgery up to 1971, and criminal damage after 1971. This then becomes a problem in matching offence codes. The first three levels of the now four level ACPO codes used on the PNC have not changed over time. These codes have been converted to Home Office (OI) offence codes using an in-house SAS conversion program (DUMP 3). However, the conversion routine converts to the current definition of the codes and no attempt is made to convert to historic codes. Thus, a conviction for forgery in 1969 would be coded 58 on the OI, but would be converted to code 60 by the conversion routine.

2. Sudden changes in coding. This has happened to the disposal codes. One set of codes was used pre-1990 and a different set of codes was used from 1990 on. Again, RDS has converted the PNC disposal codes to Home Office (OI) disposal codes using the 1990-on coding scheme. Thus any disposals before 1990 will show differences.

3. Changes in structure. The Offenders Index has recently changed format. with the police code now moving from level 2 to the level 3 record. This is in recognition of the fact that convictions from more than one police authority can be brought together in the same court room. The Offenders Index data used in this study confirms partly to the old coding scheme and partly to the new.

4.4 Recommendations regarding OI data

Our single recommendation is to provide users with a comprehensive code book containing all standard list and non-standard list offence codes. For example, one of the most popular offence codes in the Offenders Index is code 918. However, this code is not identified in the current issue of the codebook. There are many other codes omitted from the list which users find regularly. For example, the standard list motoring offences with codes of 803, 804, 809, 810, 918-923, 925,959,963, 965, 966, 971 and 998 as well as a breach offence with code 830 are all missing from the 1998 codebook.

4.5 Advantages and disadvantages of the OI

Advantages	Disadvantages
Complete conviction history for all offenders since 1963	Excludes Scotland, N. Ireland.
Information on conviction dates, offences, disposals, courts.	Not complete history for older offenders
Can search offenders index by name, by offence code, etc	Delay for information to be collected and processed
Subset of OI data (anonymised) available at ESRC data archive	Criminal histories built up by matching process
	No cautions, warnings
	No dates of offence – problem of pseudo-reconvictions

5. The problem of offence and disposal coding

We have established in Section 4.3 that the Offenders Index has had a history of changes of offence and disposal coding (with a major re-structuring in 1990). In this section we discuss this problem in more detail, as well as commenting upon the problem of converting offender and disposal codes on the PNC to the coding schemes used on the OI.

5.1 Offence coding

The yearly changes have recycled old offence codes, so that some have completely changed their meaning over time – a code for fraud later becomes criminal damage. This appears to have been driven by the need to keep offence codes below 100 as indictable offences, and those above 100 as summary offences. The documentation of Offenders Index codes is variable - offence categories known to the 1996 version of the Offenders Index codebook were lost from the 1998 version. No year 2000 version of the codebook has appeared and it appears that the work of providing codes for all disposals and all standard list offences, let alone non standard list offences, has fallen short of the mark.

We now turn to the PNC data. The Key Performance Indicator on the standard on offence codes refers to using the hierarchically structured ACPO offence codes through to their 4th level. However, examining past PNC records in any dataset will produce a mixture of ACPO codes, CCCJS codes and free text descriptions of offences. This type of data is not conducive to research; in particular to the task of matching offences to those contained within the Offenders Index.

As a way of proceeding on this research study, negotiations with RDS led to the supply of an extra data field in the PNC data files; a field that had been created by code (referred to as DUMP3) - which converted the two numeric PNC offence codings into the current version of the Home Office offence codes as used on the Offenders Index.

We have not looked in detail at the internal workings of DUMP3; but only look at how well it initially appears to do the job it is supposed to. First, we have seen that the routine generates at least one invalid offence code '-1952. A far more important failing in relation to this research is that it converts to the latest definition of the Offenders Index codes and fails to take account of changes in coding over time. This means that matching of offences between the Offenders Index and the PNC becomes more and more inaccurate as the records go further and further back in time. A well-founded matching process needs to apply the correct codes to the correct time period, and RDS need to commit resources to update this procedure. In addition, we would recommend that future codings for new offences be given entirely new codes so that the issue of recycling old codes no longer becomes an issue.

Documentation provided by RDS covering offence codes were as follows:

Offence codelist.xls	8,966 offence categories;
Dump.xls	9,118 offence categories
ACPOcode.xls	21,591 ACPO and CCCJS codes and offence categories.

The ultimate spreadsheet that RDS could provide which contained the methodology used to recode offence codes into Home Office codes contained the following

Dump3SAS.xls -approx 3,700 recodes from ACPO codes
-approx 8,700 recodes from CCCJS codes
-approx 4,000 recodes from both sources.

No spreadsheet of offences was therefore exhaustive, and re-coding of known offences was limited. For example, 14% of the offence categories (1372 offences) were recoded in dump3SAS.xls to the missing code of 9999. The108 Acts relating to these 1372 offences are listed in Appendix 5. There were a further 554 types of offences marked solely by a hyphen in the DUMP spreadsheet; their 42 Acts are given in Appendix 6.

The quality of the conversion routine supplied for offence codes is clearly not satisfactory. RDS maintain that the offences recoded 9999 are mainly non-recordable offences relating to breaches, but have not clarified whether this means that they should not appear on the PNC. In any case, many of the acts listed in Appendices 5 and 6 are major pieces of criminal legislation – the Criminal Justice Acts of 1961,1982,1987 and 1988 are examples if this.

The ideal documentation to enable decoding and comparing all e.g. post 1963 offences on the OI and the PNC would have (at least) the following columns:

- A line for each offence description that has existed since e.g. 1963.
- Date that it became possible to prosecute for this offence
- Date that it NO LONGER became possible to prosecute for this offence
- Alternatively -date that the continuance of this offence category has been verified
- ACPO 4 level code for offence (if existing at the time of the offence)
- Pseudo ACPO 4 level code for offence (i.e. a retrospective code applied along present principles)
- Home Office OI codes used and date changes
- CCCJS codes
- Criminal Statistics Offence Group
- Offenders Index offence group

It is recommended that RDS develop such documentation as an essential prerequisite to developing matching at the offence level.

5.2 Disposal codes

The situation is similar for codes for disposal types. The OI disposal codes changed in 1991, with a new three-digit coding system being introduced. Procedures also exist at the Home Office for converting PNC disposal codes to OI disposal codes, but these are very crude. Converted disposal codes were made available to us for every PNC dataset, but we did not have access to the code or spreadsheet which carried out this conversion, and so are unable to comment further. The conversion routine makes no distinction between effective and non-effective sentences and also ignores the pre-1991 coding. To proceed, suitable documentation which is similar in nature to that outlined above for offence codes would be drawn up for disposal codes, showing changes over time. It is recommended that RDS develop these over a period of time-

6. The research studies

This report uses data from five research studies. The ideal basic structure of the data for each study is as follows:

a) a *link file* with one record for each individual with some or all of the following information: surname, forenames, date of birth, sex, research ID. This file is the basic start point for the research studies. It is this file which is passed to the OI for the retrieval of criminal histories. The file should also have a PNC identifier which has been added by NIS following the procedure described in Section 3.2.

b) A dataset of *Offenders Index records*. This follows the format described in Section 4. The level 1 record for each individual will contain the research ID; this provides a link between the link file and the OI data. There needs to be at least one level 1 record on the OI for each individual in the link file.

c) A dataset of *PNC records*, following the format described in Section 3. The level 1 record for each individual will contain the PNC ID, which provides the key between the link file and the PNC file.

Figure 6.1 shows this ideal arrangement. Mr. Sillitoe, a male born on 30th April 1972, has been traced on both the OI and the PNC datasets (although with slightly different personal information). The research identifier provides the link to the OI dataset, and the PNCID the key to the PNC dataset.

The five research studies are:

a) Persistent Young Offenders, with individuals identified from PNC data

b) Probation Pathfinder offenders from Devon, Hereford and Worcester, and Teesside, with individuals identified from the local probation services.

c) Offenders used as the control group in research on the Prison Service's Home Detention Curfew scheme, with individuals identified from the prison service.

d) Offenders sentenced to any disposal in one week in 1997, as drawn from the annual OI 'Sentencing sample'

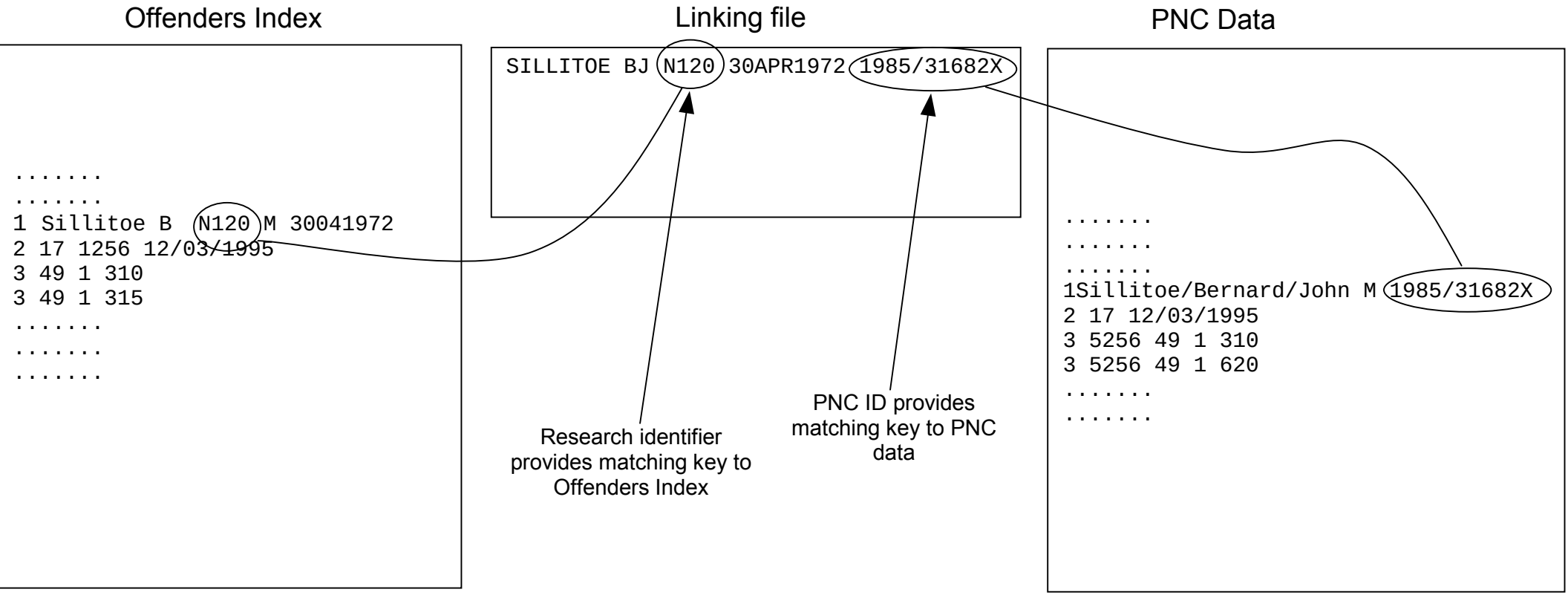
e) Offenders from the Strategic Alliance Prospective Reconviction Study relating to four Probation Service areas (with individuals identified from probation service records).

The raw data files supplied often contained duplicated entries, and in some cases the PNC and/or OI files contained records for individuals not in the link file. The files were 'cleaned' by deleting duplications and superfluous data, as detailed below. The numbers relating to PNC files refer to offence level records.

6.1 Persistent Young Offenders

The link file and PNC file each contain 8531 cases identified by PNCID. The OI file contains 8529 cases counted by PNCID. 8107 different OI numbers are present with 13 duplications (corresponding to different PNCIDs, and suggesting that the same OI criminal record has been matched to different individuals). No OI record was found for 409 individuals. Merging the files on PNCID shows that the records for 8120 individuals can be linked, however (at least) 13 of these will be linked with the wrong OI record. There is also some mismatching of sex and age between Persistent Young Offender PNC data and OI data.

Figure 6.1: The ideal linking process for individuals



6.2 Pathfinder

The Pathfinder study took place in three different probation service areas. Each area has its own characteristics and we describe them separately.

i) Teesside

Individuals in the link file were identified by 605 unique local serial numbers. 42 PNCIDs were duplicated once, 3 were duplicated twice and 40 were missing. Duplicated entries in all cases had the same name and date of birth and the duplications were deleted, retaining (where information was available) the entry corresponding to the latest sentencing date. A further duplicated individual, identified by name and CRO number was also deleted, leaving 556 individuals in the dataset, counted by serial number. After removing the duplicated records of the individuals identified in the link file from the OI file, there were 532 different serial numbers in the OI file. These corresponded to 506 different OI criminal records, with three duplications (relating to different individuals). 23 OI criminal records were missing.

The PNC file contained 522 individuals identified by PNCID. Again there were duplicate entries: 41 offender records duplicated once and 5 duplicated twice. The duplicate entries were deleted.

Merging the files indicated that of the 556 individuals in the link file, 508 could be linked to an OI record, 498 to a PNC record and 476 to both. However, some of these links will be incorrect as the wrong individual may have been identified at NIS or by the OI matching procedure. We discuss this in the next section.

ii) Devon

Individuals in the link file are identified by 179 different serial numbers. As three PNCIDs were duplicated, in each case the entry corresponding to the earlier sentencing dates was deleted. This left 176 individuals in the file, counted by serial number.

The duplicated records corresponding to the serial numbers deleted from the link file were also deleted from the OI file, leaving 184 different serial numbers in the file, including eight not present in the link file. These were deleted, leaving 176 serial numbers corresponding to 155 OI records. OI records were not present for 21 individuals. The PNC file contained 159 individuals identified by PNCID.

Merging the files indicated that of the 176 individuals in the link file, 155 could be linked to an OI record, 137 to a PNC record and 127 to both.

iii) Hereford and Worcester

The Hereford and Worcester data is interesting as it consists entirely of women offenders. This introduces the extra complexity of family name changes. Each individual in the link file was entered separately for each sentencing date, and identified by a serial number with a numerical code followed by an alphabetic suffix corresponding to the sentencing date. The file contained up to six duplicated entries per individual. When the duplications were deleted,

leaving in each case the entry with the latest sentencing date, the 731 serial numbers in the raw data were reduced to 589. Three cases were noted where the surnames (of women) had changed between sentencing dates. Only the entry corresponding to the last sentencing date was retained.

Checking on PNCID revealed five further duplications, with two referring to differently named individuals, and three to the same individuals. The three duplicated individuals were deleted as above, leaving 586 individuals in the file. The records corresponding to the duplicated individuals identified in the link file were also deleted from the raw OI data, together with two further entries, where duplication was found. The remaining 587 serial numbers could be linked to 484 OI records, with some duplication of OI numbers.

The PNC file contained 535 individuals identified by PNCID. Merging the files indicated that of the 586 individuals in the link file, 480 could be linked to an OI record, 528 to a PNC record and 437 to both. Though some discrepancy on gender was identified (probably errors in the OI file), the individuals in this file appear to be all women.

6.3 Strategic Alliance

The link file contains 1418 individuals identified by a Probation Service CRN code, gender, date of birth and PNCID and CRO numbers (the latter two with 3 duplications of each). 16 PNCIDs and 114 CRO numbers are missing.
The two PNC files supplied contained 1378 (Nov. 2000) and 1376 (June 2001) PNC records. The OI file supplied contained 1336 different OI records, in some cases linked to two different Probation Service numbers.

Merging the files indicated that of the 1418 Probation Service numbers in the link file, 1339 could be linked to an OI record, 1380 to a PNC record and 1308 to both (Nov), and 1377 to a PNC record and 1307 to both (June). In this report we have reported the analysis of only the June 2001 dataset in detail, although both datasets were analysed, with in general very similar results.

6.4 Home Detention Curfew

Dodgson (2001) describes the Home Detention Curfew (HDC) scheme which came into operation in January 1999. It allows for the release of eligible prisoners up to 60 days early on an electronically monitored curfew. The control sample used in this study was used; names were taken from the Prison Service and tracked on both the OI and PNC.

Individuals in the link file are identified by prison number. 8250 prison numbers were entered once, 26 entered 4 times and 2 entered 16 times. There are 13 instances where two different prison numbers share the same PNCID. Of the 8265 different PNCIDs in the file, 8250 were entered once, 13 were duplicated 7 times and 2 duplicated 15 times. As no names are given in the file, it is not possible to identify how many unique individuals are present. However, inspection of the OI file suggests that often when two prison numbers are linked to the same PNCID, two different individuals are involved.
All prison numbers were therefore retained, but duplicated entries were deleted, leaving 8278 different prison numbers and 8265 different PNCIDs. Thus, with records being matched for each prison number, there may be more than one matched record per individual.

The OI file supplied contained records for 8613 different prison numbers. Records for individuals not in the link file were deleted, leaving 8278 prison numbers and 7717 separate OI numbers, with 561 missing. Six duplicated OI numbers are linked to identically named individuals with different prison numbers, and often different PNCIDs.

The PNC file contains 8243 individual level records, with 15 being duplicated three times. These duplications were deleted. Merging the files indicated that of the 8278 prison numbers in the link file, 7717 could be linked to an OI record, 8254 to a PNC record and 7702 to both.

6.5 Sentencing Sample

The (version 2) link file contained 10268 different OI numbers, with 9767 different PNCIDs, of which 9734 are entered once, 31 are duplicated once and two twice. 49 PNCIDs are missing, and a further 417 are given as 'NO TRACE'. The PNC file contained 8936 different offence level records, with 32 records duplicated once and 2 duplicated twice. These duplications were deleted. The OI records for the 10268 individuals in the link file were selected from the large 1997 OI file supplied. The latest conviction date in these files is 31/12/1997. Merging the files shows that of the 10268 individuals in the link and OI files, the records for 8898 individuals can be linked to the PNC file.

7. Preparing data for matching

7.1 Introduction

We define 'hybridisation' as the process of producing a hybrid dataset, with conviction records contributing from more than one source. We distinguish between an operational system, where the intention would be to include new conviction records from the PNC onto the OI which would never have appeared on the OI – examples are Scottish convictions, convictions for non standard list offences and cautions - and this comparative work. The comparative work outlined below is concerned with how similar the two datasets are if similar criteria are used to restrict both datasets. We refer to this as the levelling of the datasets. Once the datasets have been levelled, the matching process can begin.

The procedure for preparing data for hybridisation is therefore as follows:

- to level the datasets so as to ensure a level playing field in any comparative work. For example, the OI contains mainly standard list offences; the PNC data contains all offences. The PNC will therefore contain more offences than the OI. Four levelling criteria were identified.
- to match the link file with the OI and PNC files at the level of the individual. This will produce a composite file with three sets of information – the link information, the PNC information and the OI information. This file provides information on whether the names, dates of birth, gender and other details are the same in the three files.
- within each matched individual, to match the court disposal records on the OI to those on the PNC.

This section is concerned with the levelling of the datasets; Section 8 discusses in detail the matching process.

7.2 Producing a composite file at the level of the individual

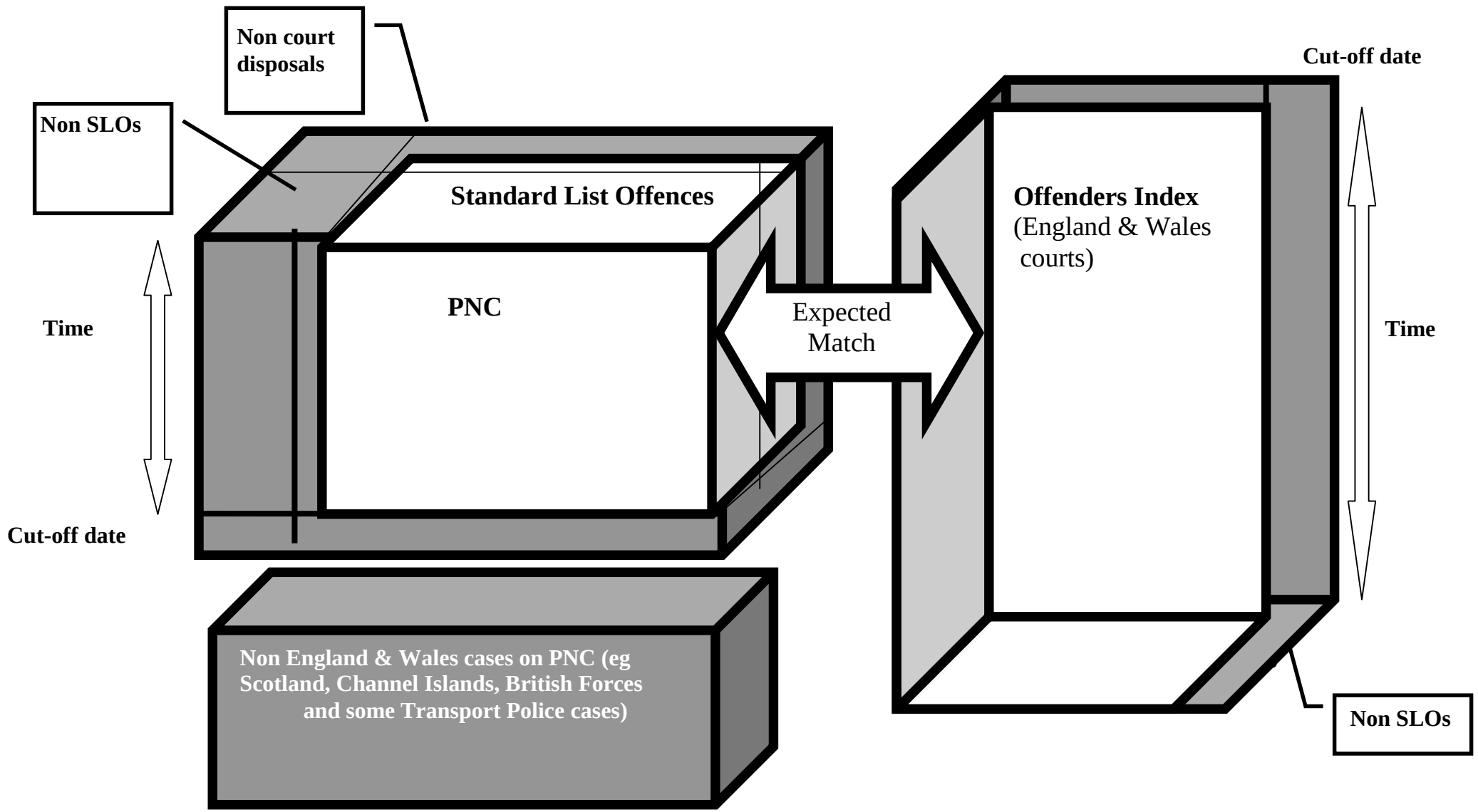
There are three tasks in this process:

- Identify and delete duplicated individuals in the link file. Some link files, particularly those originating from Probation services, have the same individual entered more than once. Where possible, the latest entry was retained, with all other duplications being deleted.
- Delete duplicated records from the raw OI and PNC data files. This aids certain statistical packages in their aggregation processes and avoids any overcounting of offences.
- Match individuals by merging the link file with the PNC file aggregated on PNCID and with the OI file using the unique individual ID.

7.3 Levelling the court-level datasets

As the PNC holds many categories of records that are not the concern of the Offenders Index, any comparison of data-completeness requires a procedure to restrict both datasets to a common level of detail. This process, termed levelling, is represented in Figure 7.1. The four levelling criteria identified were:

Figure 7.1 The levelling of the PNC and OI datasets to a common level of detail



a) levelling criterion 1 – conviction

The PNC dataset contains information on cautions, warnings, and cases awaiting trial, as well as convictions. The OI only contains convictions. All court dates and associated offences not relating to a conviction were identified with an indicator and excluded from the comparison process.

b) levelling criterion 2 – standard list offence

As stated before, the OI contains mainly standard list offences; whereas the PNC data contains all offences. Both files were brought to a level playing field by excluding all non-standard list offences from both datasets. Note that the OI needs to have non-standard list offences excluded as non-standard list offences may be present where the principal offence is standard list. No 'principal offence' is a non-standard list offence on the OI.

c) levelling criterion 3 – caution/conviction date not too recent or old for match

The datasets that we are analysing are generated by other research projects, and the requests for criminal conviction histories to PITO and to the OI were made at different times. We therefore need to take into account the date of request. Even if the date of request is the same for the PNC and OI, the OI is updated quarterly six months in arrears and is likely to contain less recent information. We account for these differences by finding the latest court record in the OI dataset, and the latest court record in the PNC, and taking the less recent of these two dates. All convictions more recent than this date would be excluded. For example, the OI sentencing sample has a most recent conviction date of 31 December 1997, the PNC date is more recent. For this dataset, all convictions from 1998 on will be excluded from the PNC file.
A separate issue relates to convictions before 1963. As the Offenders Index contains only convictions recorded in 1963 or later, and 1963 dates on the OI are unreliable, all convictions with conviction dates prior to 1964 were excluded.

d) levelling criterion 4 – sentenced within same jurisdiction

The PNC data includes data from a wide number of different jurisdictions. These additional data records consist mainly of Scottish conviction data, but it is also possible to get conviction data from the Isle of Man, Jersey, Guernsey, Northern Ireland, the Ministry of Defence, the Army, the Navy and the British Transport Police. The OI, in contrast, include only convictions to civilian police in England and Wales. The datasets were levelled by making sure that only England and Wales convictions brought by civilian police forces were included in the two files.

Details of the number of records which meet each of the levelling criteria are given in Appendix 3 for the research datasets.

8. The matching process

The matching process uses three files: a link file with data at the individual level and two files of court conviction records for the individuals in the link file, one obtained from the PNC and one from the OI.

A purpose-written FORTRAN program carried out the matching to the specification described below. The software allowed user intervention where this was necessary – both to match additional records and to identify individuals as wrongly matched.

8.1 Matching individuals

Information used in this match includes the PNC surname, forenames, sex, date of birth and the OI surname, initials, sex and date of birth. Both surnames were standardised by converting all family names starting with Mac to Mc and removing apostrophes. The forenames were converted to initials. Years of births were converted to two digits. A *provisional match* for individuals is accepted if every PNC item (surname, initials, birth date, sex) agrees with the corresponding OI item. The software then proceeds to match court dates within individuals.

Assuming that the records refer to the same individual, a list of unique court dates from both files is constructed. The percentage of occasions where the court dates are common to both files is then calculated and displayed. If a provisional match for individuals has been found, and the percentage of common court dates is 60% or above, then the individuals are taken to be the same person. If not, the user is asked to determine manually whether the conviction records belong to the same person.

At this stage, the software gives additional summary information from both the PNC file and the OI file. This consists of the maximum and minimum court identifier, the maximum and minimum police codes, a summary of the number of offences at each court date for each of the ten offence groups and the total number of offences within each court date. The latest date of offence for the PNC court date is also displayed.

The choice for the user at this stage is to **accept** the two conviction records as belonging to the same individual, to **reject** the match as two conviction records belonging to different individuals, and to **partially** accept the match as correct. Partial matches often occur with common family names such as Smith, Green, Taylor, etc. where two or more individuals with similar names and dates of birth on the Offenders Index database have been merged erroneously into a single conviction history. If this ‘partial match’ option is chosen, the user needs to specify which conviction dates belong to the target individual, and which do not. Those court dates not belonging to the target individual are jettisoned from the matched file. Note that we do not allow this possibility for the PNC data, as we have been assured that PNC data is merged only on external forensic evidence such as fingerprinting.

If we accept or partially accept the conviction records as belonging to the same individual, then the matching process proceeds to stage two.

8.2 Matching court dates within individuals.

User intervention is allowed to match together court appearances which have different dates but which appear to refer to the same court appearance. In general, these will be identified by dates differing in a single digit (mistyping error) or differing by a period of weeks, usually in the same court (but perhaps with different adult/youth court signifier) and with similar but not identical offence summary information. The software currently allows '1-1' matching of court dates; that is, once a PNC court date is matched to an OI court date, no other PNC court date can be matched to the same OI court date. There is some evidence in the study that '1-many' and 'many- 1' matching might be required. We refer to this later in the example matches.

8.3 Matching within court dates

In this report, we considered the problem of matching offence to offence within a matched court date. That is, each PNC offence would be matched to its equivalent OI offence. While this is desirable and must be a long term aim, there are some problems which need to be confronted. Any match at the offence level within a conviction would need to work on the offence code and on the disposal information. We have already identified in Section 5, the problems which arise in comparing and converting the two coding schemes. While the coding translation is adequate for matching at the court date level, it is not adequate for matching at the offence code level. We have therefore taken the view that the task of matching at the offence level should be postponed until more robust conversion routines are developed.

Once documentation is improved, it is recommended that RDS undertake a similar exercise to that undertaken here to investigate the robustness of merging data at the offence level, and to develop robust methods of merging such data. Development of such techniques will need much more in the way of manual matching as any automatic matching procedure would need a large database of human decisions on when two offences are likely to be the same and when they are not.

9. Example matches

9.1 Case 12 Pathfinder Teesside

```
*************
*           *
* Case   12 *
*           *
*************
   BRC Research ID     CRO no          dob     sex
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
LINK-  T1030/97       188764/82M                *LION          RORY
PNC -C                188764/82M   20/ 1/1965  M  *LION/RORY/STUART
OI  -                 18876482     20/ 1/1965  M  *LION              RS
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
offender information is the same

 Out of  21 conviction dates found,  80.00% ( 17) were common to both files.
```

	PNC DATA												OI DATA											
	COURTCODE		POLICE		OFFENCES				NOFF	MXOYR	D	M	Y	COURTCODE		POLICE		OFFENCES				NOFF	D	M	Y
	-	+	-	+										-	+	-	+								
1	1249	1249	17	17		1t			1	0	12	NOV	1982	1249	1249	17	17		1t			1	12	NOV	1982
2	1249	1249	17	17		1t			1	0	17	DEC	1982	1249	1249	17	17		1t			1	17	DEC	1982
3	1249	1249	17	17		1t			1	0	28	NOV	1983	1249	1249	17	17		1t			1	28	NOV	1983
4	1249	1249	17	17	1b				1	0	6	DEC	1984	1249	1249	17	17	1b				1	6	DEC	1984
5														1249	1249	17	17	1b				1	18	JUL	1985
6	1249	1249	17	17	2b				2	0	8	OCT	1985	1249	1249	17	17	2b				2	8	OCT	1985
7	1249	1249	17	17		3t			3	0	25	MAR	1986	1249	1249	17	17		3t			3	25	MAR	1986
8	2536	2536	12	12			1o		1	0	11	JUN	1986	2536	2536	12	12			1o		1	11	JUN	1986
9	1249	1249	17	17	3b				3	0	8	JUL	1986	1249	1249	17	17	3b				3	8	JUL	1986
10	9998	9998	17	17	1b				1	0	1	FEB	1988	460	460	17	17	1b				1	1	FEB	1988
11	460	460	17	17	2b	1t			3	0	6	JAN	1989	460	460	17	17	2b	1t			3	6	JAN	1989
12	460	460	17	17	5b	1t	1c	2o	9	0	8	FEB	1990	460	460	17	17	3b		1c		4	8	FEB	1990
13	9998	9998	17	17	1b				1	0	7	MAR	1995	1249	1249	17	17	1b				1	7	MAR	1995
14	1249	1249	17	17			1c		1	1997	19	MAY	1997	1249	1249	17	17			1c		1	19	MAY	1997
15	1249	1249	17	17				1o	1	1997	3	SEP	1997												
16	1249	1249	17	17	1v				1	1997	12	DEC	1997	1249	1249	17	17	1v				1	12	DEC	1997
17														1249	1249	17	17	1v				1	9	MAR	1998
18	1249	1249	17	17	1v			1o	2	1998	19	MAR	1998												
19	1249	1249	17	17	1v			1o	2	1998	17	APR	1998	1249	1249	17	17	2v				2	17	APR	1998
20	1249	1249	17	17	1v				1	1998	9	JUL	1998	1249	1249	17	17	1v				1	9	JUL	1998
21	1249	1249	17	17				1o	1	1998	2	DEC	1998	1249	1249	17	17	1v				1	2	DEC	1998

```
Any manual matches to make?
```

9.1 Case 12 Pathfinder Teesside

This gives an example of the screen output from the matching software for an unproblematic case. Following the case number, the next six lines of information show data at the individual level. Each case has three separate lines of information at the top. The first line is the information from the link file followed by lines containing the PNC information and finally the Offender Index information. Each line contains a CRO number followed by a surname and forename(s). The PNC gives the full forenames, whereas the OI only gives the initial(s). The PNC and OI also have the date of birth and the sex of the individual. In this case all the individual information agrees, i.e. the surnames match in all three fields, as do the initials, also the date of birth and the sex match for the PNC and OI fields and CRO numbers are the same for the Link, the PNC and the OI.

There then appears summary information for each conviction taken separately from the PNC files and the OI files. These are automatically matched for identical court dates and if identical dates are found they are linked. If there had been a discrepancy, e.g. in the offender details there were different dates of birth in the OI and PNC information, or if there were spelling mistakes in the name, the operator would have been asked whether they wanted to accept the information as belonging to the same person or not. They would also be able to opt for a partial match where they could choose whether to accept only some records from the OI as records in the new combined merged file.

The summary information displayed for each court date shows PNC data on the left, and OI information on the right. The information displayed starts with the maximum and minimum court code and the maximum and minimum police codes. (There are cases where the court and police codes differs within a court date). There then follows summary offence information for the ten Criminal Statistics categories. For example, '2b' means two burglaries; '1t' means a single theft offence, and 9c means 9 or more criminal damage offences. The codes used are
v (violence), s (sexual), r (robbery), b (burglary), t (theft), f (fraud), c (criminal damage), d (drugs), m (motoring) and o (other).
This is followed by the total number of offences at that conviction date (with '99' representing 99 or more) and, for the PNC data only, the latest end year of all the offences at that court date.

Where unlinked records exist on both the OI and PNC, the software will ask the operator if they would like to make any further manual matches. In this particular example line 18 (PNC) and line 17 (OI) appear very similar. The dates differ by only ten days and the police and court codes are the same. It is highly probable that it this is the same court record and so the conviction records are linked. In this case, an input error on the date may have caused the discrepancy, either on the data entry on the PNC, or for the court database.

In general, the OI and PNC histories agree in the fine detail, although court codes are sometimes missing for the PNC records (code 9998). It is instructive to see that the 'other' offence category is often used for the PNC data, which indicates a failure of the offence code conversion routine.

9.2 Case 253 Pathfinder: Hereford and Worcester

```
*************
*           *
* Case  253 *
*           *
*************
   BRC Research ID     CRO no          dob     sex
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
LINK-  9700208a       52489/89H                  *MOUSE         MINNIE
PNC -C                52489/89H    13/10/1960  F  *DUCK/MINNIE /ANN
OI  -                       0     13/10/1962  F  *MOUSE            M
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
offender information is different

 Out of   4 conviction dates found,   0.00% (  0) were common to both files.
                      PNC DATA                                                    OI DATA
      COURTCODE POLICE        OFFENCES        NOFF MXOYR D   M   Y      COURTCODE POLICE        OFFENCES        NOFF  D   M   Y
        -    +   -  +                                                    -    +   -  +
      ------------------------------------------------------------------------------------------------------------------------
   1   443  443  36 36            3f             3    0 18 DEC 1990  |
   2  9998 9998  53 53          1t               1    0 14 DEC 1994  |
   3  9998 9998  20 20          3t               3    0 28 FEB 1995  |
   4                                                                   |  2048 2048  33 33          1t               1 26 FEB 1997
      ------------------------------------------------------------------------------------------------------------------------
Accept individual records as the same person (Yes/No/Partially)?
```

This is an example where records were not accepted as belonging to the same person. The names are different (although this in itself is not always an indicator that it is a different person – especially if the offender is female), the date of birth is different and the offences occur in different areas. The PNC name differs from the link name, whereas the OI name is similar to the link name. Moreover, none of the conviction dates match. Though it is still just possible that the offender is the same person (for example, the OI might have the offender history when the offender was married, the PNC might have the history when the offender was single), with the information available it was rejected as a match.

9.3 Case 509 Sentencing Sample

```
*************
*           *
* Case  509 *
*           *
*************
   BRC Research ID     CRO no          dob     sex
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
LINK-         440649  88974/72E                  *MOUSE           M
PNC -C                88974/72E    16/ 7/1955  M  *MOUSE/MICKEY/THE
OI  -                 8897$/72E    16/ 7/  55  M  *MOUSE           M
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
offender information is the same

 Out of  43 conviction dates found,  60.00% ( 26) were common to both files.
```

	PNC DATA										OI DATA									
	COURTCODE		POLICE		OFFENCES			NOFF	MXOYR	D M Y	COURTCODE		POLICE		OFFENCES			NOFF	D M Y
	-	+	-	+							-	+	-	+					
1	6693	6693	2	2	1t			1	0	15 APR 1966									
2	6635	6635	2	2	1t			1	0	16 JUN 1966									
3	9998	9998	2	2	1t			1	0	17 JUN 1966									
4	2651	2651	2	2	1b			1	0	22 SEP 1967	6693	6693	1	1	1b			1	22 SEP 1967
5											6693	6693	1	1	1t			1	11 OCT 1968
6	9998	9998	2	2		1o		1	0	6 JUN 1969									
7	2651	2651	2	2	1t			1	0	5 OCT 1972									
8	479	479	46	46	1t			1	0	24 JAN 1973	845	845	46	46	1t			1	24 JAN 1973
9	418	418	2	2	1b	1f	1o	3	0	21 MAR 1973	853	853	1	1	1b	1t1f		3	21 MAR 1973
10	2656	2656	2	2		4f		4	0	23 MAY 1975	2656	2656	1	1		2f		2	23 MAY 1975
11	2656	2656	2	2	1t			1	0	3 JUN 1975									
12	434	434	46	46	1t			1	0	12 AUG 1975	723	723	46	46	1t			1	12 AUG 1975
13	440	440	2	2	3t3f			6	0	6 JAN 1976	850	850	1	1	4t		1m	5	6 JAN 1976
14	414	414	42	42	1t			1	0	2 APR 1976	720	720	42	42	1t		1m	2	2 APR 1976
15	2656	2656	2	2	1t			1	0	29 JUN 1976	2656	2656	1	1	1t			1	29 JUN 1976
16	413	413	2	2	1t			1	0	28 APR 1978	726	726	1	1	1t			1	28 APR 1978
17	9998	9998	2	2	1b			1	0	12 JUL 1978	850	850	1	1	1b			1	12 JUL 1978
18	440	440	2	2	1t			1	0	6 OCT 1978	850	850	1	1	1t			1	6 OCT 1978
19	2660	2660	2	2	1t			1	0	12 OCT 1978	2660	2660	1	1	1t			1	12 OCT 1978
20	440	440	2	2	1t			1	0	13 JUL 1979	850	850	1	1	1t			1	13 JUL 1979
21	2644	2644	2	2		1o		1	0	12 FEB 1981	2644	2644	1	1			1o	1	12 FEB 1981
22	440	440	2	2	1t			1	0	8 NOV 1982	850	850	1	45	2t			2	8 NOV 1982
23											2643	2643	1	1			1o	1	25 JAN 1983
24	440	440	2	2	3t8f	1o		12	0	15 MAR 1984	440	440	46	46	7t9f			16	15 MAR 1984
25											2643	2643	1	1	1t	1d		2	14 APR 1992
26	2656	2656	2	2	1t1f			2	0	15 APR 1992	2656	2656	1	1	1t1f			2	15 APR 1992
27	2643	2643	2	2	1t	2d		3	0	1 MAY 1992									
28	2656	2656	2	2		1d1o		2	0	19 MAY 1992	2656	2656	1	1	1v		1d	2	19 MAY 1992
29											2651	2651	1	1			1o	1	23 DEC 1992
30	2643	2643	2	2	4t3f			7	0	12 MAY 1993	2643	2643	1	1	4t3f			7	12 MAY 1993
31	2727	2727	2	2	1t			1	0	3 SEP 1993									

9.3 Case 509 Sentencing Sample (continued)

This case illustrates the problem with differences in police and court codes. All the offender details, i.e. the names, the dates of birth, the sex and the CRO numbers appear to match and there is a 60% conviction date match. However, closer analysis of the court codes suggests that it may not be the same person. As the police areas also appear different this would appear to confirm that it is not the same person. Yet, if one were to look in the Offenders Index handbook one would find that there is no police area 2 (working through the data has shown that this appears to be a code that is sometimes input instead of police area 1 -the Metropolitan area). Also many of the offence details correspond as do the court disposal dates. Line 13 presents particular difficulties. If this had been a partial match case, it would have been difficult to decide whether to accept it or not as the offences do not appear to correspond, i.e. it has different police areas, different courts and different offence types. Later information, from line 24 onwards, appears to be more reliable with regards court information. The criminal histories were accepted as belonging to the same person.

9.4 Case 1445 HDC

```
*************
*           *
* Case 1445 *
*           *
*************
   BRC Research ID     CRO no          dob     sex
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
LINK-  BA007469       12463/88V                 *
PNC -C                12463/88V    28/ 3/1974  M  *BUILDER/BOB
OI  -                 12463/88V    28/ 3/1974  M  *BUILDER            B
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
offender information is the same

 Out of  16 conviction dates found,  81.00% ( 13) were common to both files.
                        PNC DATA                                                  OI DATA
      COURTCODE POLICE       OFFENCES        NOFF MXOYR D   M   Y      COURTCODE POLICE        OFFENCES        NOFF  D   M   Y
        -    +   -  +                                                    -    +   -  +
      ---------------------------------------------------------------------------------------------------------------------------
   1  9998 9998  16 16          1t              1    0 27 MAY 1987  |  5933 5933  16 16          1t              1 27 MAY 1987
   2  9998 9998  16 16      1b          1o      2    0 14 MAR 1988  |  5933 5933  16 16      1b  1t              2 14 MAR 1988
   3                                                                |  5933 5933  16 16      6b  8t  1c  1o     16 23 MAR 1990
   4  9998 9998  16 16      7b  9t  1c  1o     18    0  7 SEP 1990  |   403  403  16 16      2b  1t              3  7 SEP 1990
   5  9998 9998  16 16          2t  1c          3    0  2 NOV 1990  |
   6  9998 9998  16 16  1v  2b  4t  2c          9    0  1 MAY 1991  |  5933 5933  16 16  1v  2b  4t  2c          9  1 MAY 1991
   7   403  766  16 16  2v  5b  9t  3c         21    0  2 DEC 1991  |   766  766  16 16      5b  9t  2c         16  2 DEC 1991
   8  9998 9998  16 16          4t              4    0  9 DEC 1991  |  1933 1933  16 16          4t              4  9 DEC 1991
   9   403  403  16 16  1v  3b  9t  1c  1o     20    0  7 MAY 1992  |   403  403  16 16  2v  5b  9t  2c         25  7 MAY 1992
  10   403  766  16 16  3v  3b                  6 1994 25 FEB 1994  |   403  403  16 16  1v  3b  1t        2m    7 25 FEB 1994
  11  1933 1933  16 16          1t              1 1995 20 MAR 1996  |  1933 1933  16 16          1t        3m    4 20 MAR 1996
  12   403  403  16 16      2b                  2 1996 26 MAR 1997  |   403  403  16 16      2b                  2 26 MAR 1997
  13  1933 1933  16 16          2t      1o3m    6 1998 10 JUL 1998  |  1933 1933  16 16          2t              2 10 JUL 1998
  14  1933 1933  16 16                  1o      1 1998 19 FEB 1999  |  1933 1933  16 16                  3o      3 19 FEB 1999
  15  1933 1933  16 16          1t      1o      2 1999  9 MAR 1999  |
  16   766  766  16 16          1t      1o3m    5 1999 13 MAY 1999  |   403  403  16 16          2t      1o5m    8 13 MAY 1999
      -------------------------------------------------------------------------------------------------------------------------
Any manual matches to make?
```

This particular case shows an example of where two court dates on the Offenders Index appear to match a single court date on the PNC. Lines 3 and 4 of the OI, taken together appear to match line 4 on the PNC, with the sum of the offences on these OI records almost agreeing with the offence information on the PNC. The software, however, has automatically matched the lines according to date, thus showing a discrepancy in the offence details. The software does not currently allow one-many matching, and the OI dates cannot be combined.

9.5 Case1146 HDC

```
*************
*           *
* Case 1146 *
*           *
*************
   BRC Research ID     CRO no          dob     sex
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
LINK-  AV009565       41106/93L                 *
PNC -C                41106/93L     1/ 1/1975  M  *WILLIAMS//STUART/CLIVE
OI  -                 41106/        1/ 1/1935  M  *WILLIAMS               S
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
offender information is different

 Out of  77 conviction dates found,  12.00% ( 10) were common to both files.
```

	PNC DATA										OI DATA									
	COURTCODE		POLICE			OFFENCES		NOFF	MXOYR	D M Y	COURTCODE		POLICE			OFFENCES		NOFF	D M Y
	-	+	-	+							-	+	-	+						
1											2918	2918	20	20	2s		1o	3	1 APR 1974
2											2908	2908	20	20		1c		1	14 APR 1975
3											6978	6978	13	13		1t		1	15 JUN 1982
4											840	840	1	1			1d	1	27 APR 1983
5											2650	2650	1	1			1o	1	30 MAY 1984
6											2908	2908	20	20		1t		1	28 JAN 1985
7											5940	5940	16	16		1t		1	22 OCT 1985
8											2908	2908	20	20			9o	12	25 NOV 1988
9											413	413	1	1		2c	1o	3	13 JAN 1989
10											2908	2908	20	20			8o	8	26 MAY 1989
11											2908	2908	20	20			4o	4	2 JUN 1989
12											2651	2651	1	1		1c		1	12 JUN 1990
13											2908	2908	20	20			9o	18	22 JUN 1990
14											2910	2910	20	20			1o	1	15 AUG 1990
15											2908	2908	20	20			9o	10	9 AUG 1991
16											2908	2908	20	20			3o	3	13 MAR 1992
17											2908	2908	20	20	1v			1	24 MAR 1992
18											471	471	1	1			2d	2	11 AUG 1992
19											2762	2762	1	1			5o	5	1 SEP 1992
20											1443	1443	36	36			5o	5	12 OCT 1992
21											5734	5734	6	6	1v	1t	1o	3	21 MAY 1993
22											2908	2908	20	20			2o	2	13 AUG 1993
23											449	449	43	43			1o	1	27 AUG 1993
24											2766	2766	1	1			1o	1	9 FEB 1994
25											2650	2650	1	1	1v			1	8 JUN 1994
26	2766	2766	1	1		1o		1	0	27 JUN 1994	2766	2766	1	1			1o	1	27 JUN 1994
27	2734	2734	1	1	1t	2o		3	0	25 AUG 1994	2723	2723	1	1		1t	2o	3	25 AUG 1994
28											427	427	1	1	1v			1	3 OCT 1994
29											2650	2650	1	1			2m	2	9 JAN 1995
30											469	469	1	1		1f		1	20 JAN 1995
35	2760	2760	1	1	1t			1	1995	9 NOV 1995	2760	2760	1	1		1t		1	9 NOV 1995

```
 36                                                                    |  2646 2646   1  1            1f             1  5 DEC 1995
 37                                                                    |  1734 1734   6  6  1v                       1 15 JAN 1996
 38                                                                    |  2769 2769   1  1                1d         1 16 JAN 1996
 39                                                                    |  2908 2908  20 20                    2m     2 18 JAN 1996
 40                                                                    |  2650 2650   1  1          1t                1 24 JAN 1996
 41                                                                    |  2641 2641   1  1    6s                      6  1 MAR 1996
 42                                                                    |  2741 2741   1  1                    1m     1  6 MAR 1996
 43                                                                    |  2650 2650   1  1              1c            1 11 MAR 1996
 44                                                                    |   469  469   1  1            1f              1 22 MAR 1996
 45                                                                    |  2815 2815   1  1                       2u   2 27 MAR 1996
 46  2734 2734   1  1          1t               1 1996 16 APR 1996     |
 47                                                                    |  2650 2650   1  1                   6o       6 16 MAY 1996
 48                                                                    |  2650 2650   1  1                     2m     2  1 JUL 1996
 49                                                                    |  2641 2641   1  1    5s                       5 30 AUG 1996
 50  2734 2734   1  1          2t               2 1996 15 NOV 1996     |  2734 2734   1  1          2t        3m     5 15 NOV 1996
 51                                                                    |  2650 2650   1  1                   1o       1 30 NOV 1996
 52  2734 2734   1  1          1t               1 1996  8 JAN 1997     |
 53                                                                    |  2908 2908  20 20          1t      2o2m     5 14 JAN 1997
 54  2734 2734   1  1          1t               1 1996 30 JAN 1997     |
 55                                                                    |  2663 2663   1  1          1t        1m     2 13 MAR 1997
 56  2734 2734   1  1          1t        1m     2 1997  4 JUN 1997     |  2734 2734   1  1          1t               1  4 JUN 1997
 57                                                                    |   449  449  43 43  1v                       1 11 JUL 1997
 58  2734 2734   1  1  1v               1m     2 1997 28 JUL 1997     |  2723 2723   1  1          1t        3m     4 28 JUL 1997
 59                                                                    |  2994 2994  13 13                  9o      10 10 OCT 1997
 60                                                                    |  2656 2656   1  1                  2o       2 28 OCT 1997
 61  2734 2734   1  1          2t        1m     3 1997  1 DEC 1997     |  2723 2723   1  1          2t        4m     6  1 DEC 1997
 62                                                                    |  2650 2650   1  1                1d         1 30 DEC 1997
 63                                                                    |  2267 2267   5  5  1v                       1  7 FEB 1998
 64                                                                    |  2908 2908  20 20                  1o       1 15 APR 1998
 65  2734 2734   1  1          1t               1 1998  4 JUN 1998     |
 66  2723 2723   1  1          1t               1 1998 15 JUL 1998     |  2734 2734   1  1          1t               1 15 JUL 1998
 67  2723 2723   1  1          1t        1m     2 1998 27 JUL 1998     |  2723 2723   1  1          1t        2m     3 27 JUL 1998
 68                                                                    |  2323 2323  34 34                  5o       5  5 OCT 1998
 69                                                                    |  1733 1733   6  6                  1o       1 29 OCT 1998
 70  2734 2734   1  1          1t               1 1999  1 FEB 1999     |  2723 2723   1  1          1t               1  1 FEB 1999
 71                                                                    |   453  453   1  1  1v                       1  5 FEB 1999
 72                                                                    |  2994 2994  13 13            1f      1m     2  5 JAN 2000
 73                                                                    |  2660 2660   1  1                  1o       1 30 MAR 2000
 74                                                                    |  2908 2908  20 20                1d         1  5 JUN 2000
 75  2742 2742   1  1          1t               1 1998 19 JUL 2000     |
 76                                                                    |   427  427   1  1          1t               1 21 JUL 2000
 77                                                                    |  2978 2978  13 13                  1o       1 24 AUG 2000
    ------------------------------------------------------------------------------------------------------------------------------------
Accept individual records as the same person (Yes/No/Partially)?
```

One of the major problems with matching is that of partial matching. This occurs when two or more Offenders Index criminal histories have been brought together erroneously. This often occurs when the offenders have the same names (which are likely to be common surnames such as Green, Williams etc) and a similar date of birth. (Although the dates of birth are similar here, there is a forty year difference). Lines 4 and 5 show offences committed in the Metropolitan area. As the younger person named was only 9 years old at the time it is unlikely he has been convicted of an offence. Because the two individuals appear to be offending in similar areas and the offence types are similar it becomes extremely difficult to split the records. It is also possible that the person named by the PNC has more than one PNCID - this would explain the discrepancies between the OI and the PNC. Sometimes the case is easier to split, e.g. if one offender is prolifically offending in Cumbria and the other in Cornwall, then it is easier to identify the two separate individuals.

9.6 Case 1 PYO

```
*************
*           *
* Case    1 *
*           *
*************
   BRC Research ID     CRO no          dob     sex
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
LINK-  1993/274812Y    8032693                  *POSTMAN          P     P
PNC -C                80326/93R    10/ 5/1979  M  *POSTMAN/PAT/POSTIE
OI  -                    8032     10/ 5/1979  M  *POSTMAN              PP
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
offender information is the same

 Out of  18 conviction dates found,  72.00% ( 13) were common to both files.
                      PNC DATA                                                    OI DATA
      COURTCODE POLICE        OFFENCES        NOFF MXOYR D   M   Y      COURTCODE POLICE        OFFENCES        NOFF  D   M   Y
        -    +   -  +                                                    -    +   -  +
      ------------------------------------------------------------------------------------------------------------------------
   1  9998 9998  46 46  3v                      3    0 19 OCT 1993  |  5961 5961  46 46  1v                      1 19 OCT 1993
   2                                                                 |  5961 5961  46 46  4v                      4  4 FEB 1995
   3  5961 9998  46 46  6v      6t             12 1994 14 FEB 1995  |  1961 1961  46 46  1v      4t        2m     7 14 FEB 1995
   4  1961 5961  46 46  1v      3t        1m    5 1995 11 JUL 1995  |  5961 5961  46 46          1t              1 11 JUL 1995
   5  1961 1961  46 46  1v        1f            2 1995 22 AUG 1995  |  5961 5961  46 46          1t1f      2m    4 22 AUG 1995
   6                                                                 |  5961 5961  46 46          3t        3m    6  2 JAN 1996
   7  5961 5961  46 46  1v      3t1f      2m    7 1996  5 NOV 1996  |  5961 5961  46 46  1v      2t1f      4m    8  5 NOV 1996
   8                                                                 |  5961 5961  46 46          1t              1  2 JAN 1997
   9  5961 5961  46 46          4t        2m    6 1997 20 MAY 1997  |  5961 5961  46 46          3t        2m    5 20 MAY 1997
  10  5961 5961  46 46  1v                      1 1997 29 AUG 1997  |  1961 1961  46 46          1t              1 29 AUG 1997
  11  1961 1961  46 46          2t        1m    3 1997  2 OCT 1997  |  1961 1961  46 46          1t        2m    3  2 OCT 1997
  12  1961 1961  46 46          3t        2m    5 1998  1 MAY 1998  |  1961 1961  43 43          2t        2m    4  1 MAY 1998
  13  1626 1626  42 42          1t              1 1997 22 JUL 1998  |  1626 1626  42 42          1t              1 22 JUL 1998
  14   461  461  42 42                  1o      1 1997  4 AUG 1998  |   461  461  42 42                  1o      1  4 AUG 1998
  15  1961 1961  46 46          1t              1 1998  7 AUG 1998  |  1961 1961  46 46          1t              1  7 AUG 1998
  16  1961 1961  46 46  1v      2t      1o3m    7 1998 27 OCT 1998  |
  17  1961 1961  46 46          3t        1m    4 1999  4 JUN 1999  |  1961 1961  46 46          1t              1  4 JUN 1999
  18                                                                |   434  434  46 46          4t        2m    6  8 OCT 1999
      ------------------------------------------------------------------------------------------------------------------------
Any manual matches to make?
```

This case illustrates several problems. Firstly it shows how offence information can differ between the PNC and the OI, e.g. line 3 shows the PNC indicating the person has committed 6 violent offences and 6 thefts. The corresponding entry for the OI says the same person on the same day received a disposal for 1 violent offence, 4 thefts and 2 motoring offences there are other examples of this within this case.

Lines 17 and 18 also show another discrepancy that appeared when cases were not automatically matching. Sometimes there was an entry showing at a Magistrates court in June and what appeared to be a matching entry did not occur at the Crown Court until a few months later. This may have been because of Summer recesses –there also seemed to be gaps between November and February.

10. The results from the datasets

10.1 The matching process

The matching process was carried out on each of the five studies. For two of the studies (HDC and Sentencing Sample) there was time only to attempt to match a proportion of the sample. This was in part due to extremely late delivery of the datasets from the Home Office and in part due to continuing problems with the formats of the datasets supplied (which differed from documented format) which continued up to the final report date. The remaining datasets were matched in full. However, for the Strategic Alliance datasets, formatting problems still exist and the results should be taken only as tentative. For the HDC study, 3000 cases were matched (36% of the sample), and for the sentencing sample 4000 cases (39% of the sample) were matched. The remaining studies were matched in their entirety.

10.2 Matching at the individual level

Table 10.1 gives, for each study, details on how many individual records were available (columns 2 to 5), how many individuals with both PNC and OI data present were matched (columns 6 to 8), and the percentage of individuals matched out of those present in the original link files (column 9).

The studies varied tremendously in the degree of matching at the individual level. On the two large studies with an external data source (HDC and Strategic Alliance) around 93% of names on the link file were matched on both the PNC and the OI. However in around 2 per cent of the link file cases, our work found that the search procedure found different individuals, giving an corrected match rate of around 91%. The Pathfinder studies also had external files. Here the matching of individuals was a lot worse. After correction, 83% of Teesside cases were matched on both data sources, and this dropped to under 70% for both Devon and Hereford and Worcester. The low rate for Hereford and Worcester might be expected as the sample is all female, and it is more problematic to trace women in the criminal justice databases. However, the poor match rate for the Devon Pathfinder study remains an anomaly.

The PYO study extracted names from the PNC and searched for them on the OI – a very high match rate of over 95% was achieved, and only 0.2% were identified as incorrect. The Sentencing sample, in contrast, had a relatively poor match rate of around 88%, and a surprisingly high 3.5% of names were in fact wrongly matched. This is probably due to poor preparation of the sentencing sample link file by the Home Office –which appears to have given wrong PNCIDs for some of the names on the link file.

We might surmise that some of the PNC cases which are wrongly matched may be relatives or even twin siblings, with similar names and dates of birth. Of course, for common family names, there will be potentially many individuals with the name (such as B Jones) born on the same day and about a third of these will have criminal records. It is not surprising that incorrect matches are found, and confirmation of the match should always be through the criminal history.

Partial matches account for just under one per cent of matched records. It is recommended that a review of this problem on the OI be introduced, with the intention of systematically demerging such records as a long term aim.

Table 10.1: Matching individuals

	Number of individuals								
	Records for processing					Records for matching[3]			$\frac{(7)+(8)}{(1)}$
Study	1 In link file	2 with PNC missing-OI present	3 with OI missing-PNC present	4 with both PNC and OI missing	5 with both PNC and OI present	6 PNC and OI present - match rejected	7 PNC and OI present - partial match	8 PNC and OI present - match accepted	9 % of link file matched
1. PYO	8531	-	411 (4.8%)	-	8120 (95.2%)	20 (0.2%)	36 (0.4%)	8064 (99.3%)	94.9%
2 Pathfinder									
Teesside	556	32 (5.8%)	21 (3.8%)	27 (4.9%)	476 (85.6%)	11 (2.3%)	6 (1.3%)	459 (96.4%)	83.6%
Devon	176	28 (15.9%)	10 (5.7%)	11 (6.3%)	127 (72.2%)	2 (1.6%)	2 (1.6%)	123 (96.9%)	71.0%
Hereford and Worcester	586	43 (7.3%)	91 (15.5%)	15 (2.6%)	437 (74.6%)	44 (10.1%)	0	393 (89.9%)	67.1%
3. Strategic Alliance:	1418	32 (2.3%)	70 (4.9%)	9 (0.6%)	1307 (92.2%)	49 (3.7%)	23 (1.8%)	1235 (94.5%)	88.7%
4. HDC	3000	3 (0.1%)	180 (6.0%)	1 (0.03%)	2816 (93.9%)	40 (1.4%)	39 (1.4%)	2737 (97.2%)	92.6%
5. Sentencing sample	4000	469 (11.7%)	-	-	3531 (88.3%)	125 (3.5%)	12 (0.3%)	3394 (96.1%)	85.2%
ALL STUDIES	18267	6.2%[1]	5.5%[2]	1.1%[1,2]	92.0%	1.7%	0.7%	97.6%	90.5%

[1] Excluding PYO study
[2] Excluding Sentencing sample study
[3] Percentage base is column 5.

10.3 Matching at the court date level

For the individuals where a match or a partial match has been found, we proceed to the matching of OI court dates to PNC court dates. Court dates can either be matched automatically or manually; the manual matches are particularly of interest as they provide some insight into how matching of court dates might be detected automatically.

At the end of the process, the combined file contains court dates with originate from the PNC but not from the OI, court dates which originate from the OI but not from the PNC, and court dates which originate from both. The total number of court dates is calculated by summing all three sources. Table 10.2 below gives the results of this matching. Depending on the study, between 65% and 80% of court records appear in both data sources for matched individuals, although manual matches add between another 1% to 2% to these figures, giving a combined match rate of between

68% and 81%. There is one outlier: the low figure of 60% of matches for Hereford and Worcester again recognises the special problem of women offender histories. For this study, the PNC is contributing 30% of records to the combined data file which do not appear on the OI – it appears that OI information is particularly weak for women as many of these court dates should be present on the OI but are not. For most of the other studies this pattern continues to hold, with the PNC contributing more unmatched records than the OI. However, for the PYO study, the pattern is reversed- and the OI is contributing more records than the PNC.

Table 10.2: Sources of court date information in hybridised file, by study.

Study	Court dates				
	from PNC only	From OI only	from both PNC and OI (exact)	from both PNC and OI (manual match)	total number of court dates
1. PYO	14.6% 11,967	15.9% 13,013	68.0% 55,806	1.5% 1,230	82,016
2 Pathfinder					
Teesside	13.3% 930	5.8% 407	79.6% 5,561	1.2% 87	6,985
Devon	13.4% 210	10.1% 159	75.0% 1,178	1.5% 24	1,571
Hereford and Worcester	30.7% 529	7.6% 131	60.1% 1,034	1.6% 27	1,721
3. Strategic Alliance:	16.1% 1,518	10.9% 1,024	72.3% 6,821	0.7% 66	9,429
4. HDC	15.7% 5,648	9.9% 3,560	72.3% 26,087	2.2% 777	36,072
5. Sentencing sample	19.2% 7,866	12.6% 5,146	66.1% 27,070	2.1% 867	40,949
ALL STUDIES	16.0% 28,668	13.1% 23,440	69.1% 123,557	1.7% 3,078	178,743

The evidence from Hereford and Worcester suggests that the degree of matching at the court date level should be worse for females compared to males. However, this might be an artefact of the Hereford and Worcester dataset, and not a feature of women offenders in other studies. We examined the match rate for court dates separately for males and females, and the results are presented in Table 10.3.

All of the large sample studies show significant differences between male and female court date match rates, with the female match rate between 3% and 10% lower than the male match rate. However, Devon was an exception, and for this study, the match rate for females was slightly higher than for males. It is worth pointing out that the low match rate for Hereford and Worcester of 61.7% was echoed in the female match rates for the PYO study (61.8%) and the sentencing sample (61.3%).

Table 10.3: Match rates by gender

Study	Males		Females		
	no. matched /total	%	no. matched total	%	chi-squared test of differences between proportions; p-value
1. PYO	53937/76999	70.0%	3099/5017	61.8%	152 on 1df; p<0.001
2 Pathfinder					
Teesside	5280/6518	81.0%	368/467	78.8%	1.37 on 1df; p=0.24
Devon	1121/1470	76.3%	81/101	80.2%	0.82 on 1df; p=0.37
Hereford and Worcester	(no males)	-	1061/1721	61.7%	-
3. Strategic Alliance:	6233/8407	74.1%	654/1022	64.0%	47.7 on 1 df; p<0.001
4. HDC	25884/34703	74.6%	972/1358	71.6%	6.2 on 1df; p=0.013
5. Sentencing sample	26433/38494	68.7%	1503/2452	61.3%	57.8 on 1df; p<0.001
ALL STUDIES	118,888/166,591	71.4%	7,738/12,138	63.8%	317.6 on 1 df;p<0.001

We now examine whether the match rate appears to change according to the year of court disposal. The results are given in Table 10.4.

Table 10.4: Match rates for court dates by year of court disposal

	Before 1970	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	overall
1. PYO					4/32[a] 12.5%	5569/10600 52.5%	51463/71384 72.1%	57036/82016 69.5%
2 Pathfinder								
Teesside	29/41 70.7%	52/60 86.7%	95/111 85.6%	298/356 83.7%	651/813 80.1%	1517/1796 84.5%	3006/3808 78.9%	5648/6985 80.9%
Devon	17/18 94.4%	12/23 52.2%	32/43 74.4%	97/134 72.4%	157/226 69.5%	330/415 79.5%	557/712 78.2%	1202/1571 76.5%
Hereford and Worcester	1/5 20.0%	3/9 33.3%	15/37 40.5%	47/90 52.2%	80/156 51.3%	271/416 65.1%	644/1008 63.9%	1061/1721 61.7%
3. Strategic Alliance:	61/86 70.9%	124/177 70.1%	264/388 68.0%	555/772 71.9%	763/1020 74.8%	1317/1745 75.5%	3803/5241[c] 72.6%	6887/9429 73%
4. HDC	174/299 58.2%	443/618 71.7%	961/1390 69.1%	2001/2693 74.3%	3233/4290 75.4%	6609/8828 74.9%	13443/17954[c] 74.9%	26864/36072 74.5%
5. Sentencing sample	275/659 41.7%	729/1405 51.9%	1472/2729 53.9%	3293/5446 60.5%	5348/8098 66.0%	8371/11707 71.5%	8449/10905[b] 77.5%	27937/40949 68.2%
ALL STUDIES	557/1108 50.3%	1363/2292 59.5%	2839/4698 60.4%	6291/9491 66.3%	10236/14635 69.9%	23984/35507 67.5%	81365/111012 73.3%	126635/178743 70.8%

[a] For the PYO study, these figures relates to all disposal dates before 1990.
[b] For the Sentencing sample, these figures relate to 1995-1997 only.
[c] For Strategic Alliance and HDC figures relate to 1995-2000.

For most of the studies, the match rate gradually improves over time, with the highest match rates given in the most recent time periods - either in the 1990-4 year group or the 1995-1999 group. However, for the Teesside Pathfinder study, a different pattern can be observed, where the degree of matching starts from a low base, then quickly reaches a peak in 1970-1974, before declining slowly. The matching figure for 1998 is consistent with the previous five year

period; the decline in 1995-1999 is entirely due to the poor matching in 1999.

One of the questions which arises from the above table is whether the poorer matching in the earlier years appears to be due to disposal dates being missing from the OI or dates missing from PNC. We examine this issue in the Table 10.5. It is difficult to see any consistent pattern. However, it appears that for most of the studies, the OI and the PNC were approximately equal in the number of unmatched records they contributed to the combined dataset. However, in recent years, it appears that the PNC is contributing between 50% and 100% more court dates than the OI. However, this is not true for the sentencing sample.

Table 10.5: Match rates and OI unmatched and PNC unmatched rates for court dates by year of court disposal

		Before 1970	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	overall
1. PYO	M					12.5%[a]	52.5%	72.1%	69.5%
	O					87.5%	38.5%	12.5%	15.9%
	P					0.0%	9.0%	15.4%	14.6%
	n					(32)	(10,600)	(71,384)	(82,016)
2. Pathfinder :									
Teesside	M	70.7%	86.7%	85.6%	83.7%	80.0%	84.5%	78.9%	80.9%
	O	12.2%	8.3%	7.2%	10.1%	12.2%	6.2%	3.8%	5.8%
	P	17.1%	5.0%	7.2%	6.2%	7.7%	9.4%	17.3%	13.3%
	n	(41)	(60)	(111)	(356)	(813)	(1,796)	(3,808)	(6,985)
Devon	M	94.5%	52.2%	74.4%	72.3%	69.5%	79.5%	78.2%	76.5%
	O	5.5%	39.1%	18.6%	13.4%	15.5%	9.2%	7.0%	10.1%
	P	0.0%	8.7%	7.0%	14.2%	15.0%	11.3%	14.7%	13.4%
	n	(18)	(23)	(43)	(134)	(226)	(415)	(712)	(1,571)
Hereford and Worcester	M	20.0%	33.3%	40.5%	52.2%	51.3%	65.2%	63.9%	61.7%
	O	40.0%	11.1%	13.5%	11.1%	10.3%	6.0%	7.1%	7.6%
	P	40.0%	55.6%	45.9%	36.7%	38.5%	28.8%	29.0%	30.7%
	n	(5)	(9)	(37)	(90)	(156)	(416)	(1,008)	(1,721)
3. Strategic Alliance:	M	71.0%	70.1%	68.1%	71.9%	74.8%	75.4%	72.6%[c]	73.0%
	O	9.3%	10.2%	15.5%	13.0%	9.7%	8.7%	11.2%	10.9%
	P	19.8%	19.8%	16.5%	15.2%	15.5%	15.8%	16.2%	16.1%
	n	(86)	(177)	(388)	(772)	(1,020)	(1,745)	(5,241)	(9,429)
4. HDC	M	58.2%	71.7%	69.2%	74.4%	75.4%	74.9%	74.9% [c]	74.5%
	O	14.7%	10.7%	15.2%	14.1%	10.4%	9.8%	8.6%	9.9%
	P	27.1%	17.6%	15.7%	11.6%	14.25	15.3%	16.5%	15.7%
	n	(299)	(618)	(1,390)	(2,693)	(4,290)	(8,828)	(17,954)	(36,072)
5. Sentencing sample	M	41.7%	51.9%	53.9%	60.5%	66.0%	71.5%	77.5%[b]	68.2%
	O	16.1%	16.3%	17.0%	14.7%	9.8%	11.6%	12.8%	12.6%
	P	42.2%	31.8%	29.1%	24.9%	24.2%	16.9%	9.7%	19.2%
	n	(659)	(1,405)	(2,729)	(5,446)	(8,098)	(11,707)	(10,905)	(40,949)

[a] For the PYO study, these figures relates to all disposal dates before 1990. [b] For the sentencing sample, these figures relate to 1995-1997 only. [c] For Strategic Alliance and HDC figures relate to 1995-2000.

Key: M: Matched disposal date;
O: OI contributed disposal date, no PNC match;
P: PNC contributed disposal date, no OI match
n No. of court dates

Table 10.6 examines matching according to the PNC's back record conversion indicator. The results in general hold no surprises. BRC indicators C, E, F and O are similar in the degree of matching. The only low degree of matching in general is for BRC equal to N. However, PYO produces a surprisingly high match rate for the N category of 82.6%

Table 10.6: Match rates by BRC indicator

Study	Back Record Conversion indicator					TOTAL
	C PNC2 Convictions level only	E Enhanced data	F Full policing data	N No convictions on PNC	O Created post-migration	
1. PYO	70.4%	81.3%	68.1%	82.6%	67.4%	69.5%
2 Pathfinder						
Teesside	81.3%		78.9%	38.5%	75.7%	80.9%
Devon	78.8%		81.8%	23.9%	58.6%	76.5%
Hereford and Worcester	60.1%		70.0%	28.6%	68.1%	61.7%
3. Strategic Alliance	72.3%		68.0%	18.8%	68.4%	71.3%
4. HDC	75.1%	60.4%	77.7%	23.8%	72.4%	74.5%
5. Sentencing sample	70.0%	78.6%	67.5%	17.8%	46.1%	68.2%
ALL STUDIES	72.0%	70.5%	69.3%	21.0%	68.2%	70.8%

From this point on, we now consider only the three substantially sized studies- PYO, sentencing sample and HDC.

First, we examined the effect of police authority on the matching of disposal dates. For this analysis, we restricted the analysis to disposal dates after 1974 (previous to this date, the police authority definitions were in a state of major change) and to disposals where only a single police authority contributed to the conviction (it is possible for different police authorities to bring proceedings at the same court proceedings).

Table 10.7 gives the results of the analysis. Low match rates were consistently observed for the Metropolitan Police area, in the City of London and for some other southern counties, notably Kent, Essex and Wiltshire. High rates approaching 90% were observed for many Northern, Midlands and South Western forces, although the results were not consistent over studies.

It is important to emphasise that although the study is tabulating by police authority, a low match rate does not necessarily indicate problems with the PNC data collection – the problem might lie equally with the Offenders Index. Indeed, regarding the low match rate for the Metropolitan Police, the Offenders Index User Guide states that:

> *The (Offenders) Index is based on court proceedings returns made by the police. We are aware that there was a shortfall in data from the Metropolitan Police from 1987 to 31 July 1992. The shortfall is estimated at around 15% for 1987 to 1991 and 11% for 1992 for those sentenced for indictable offences (which form the major part of standard list offences). London-based studies are not recommended for the years in question, because the level of matching and length of criminal history will*

be affected by the shortfall. Studies which compare different areas of the country need to be interpreted with care if London is one of the areas.

Table 10.7: Match rates by Police Authority

Police Authority	Study name			Over the three datasets %
	PYO	Sentencing sample	HDC	
1,2 Metropolitan Police	66.9%	64.7%	67.4%	66.3%
3 Cumbria	85.4%	83.5%	83.9%	84.7%
4 Lancashire	83.3%	78.8%	83.2%	81.7%
5 Merseyside	80.7%	81.0%	83.1%	81.3%
6 Greater Manchester	80.0%	81.2%	84.4%	81.6%
7 Cheshire	84.5%	83.9%	88.8%	85.4%
10 Northumbria	90.9%	84.8%	85.6%	88.7%
11 Durham	83.7%	82.9%	90.5%	85.5%
12 North Yorkshire	86.0%	79.3%	84.1%	84.0%
13 West Yorkshire	81.8%	81.9%	80.0%	81.7%
14 South Yorkshire	85.0%	81.4%	86.6%	84.4%
16 Humberside	84.0%	81.2%	88.2%	85.2%
17 Cleveland	89.7%	86.5%	91.5%	89.8%
20 West Midlands	77.9%	78.0%	83.1%	79.4%
21 Staffordshire	85.3%	77.8%	83.5%	83.3%
22 West Mercia	82.9%	80.3%	80.2%	81.5%
23 Warwickshire	85.9%	85.6%	86.7%	86.1%
30 Derbyshire	88.5%	85.0%	84.3%	86.7%
31 Nottinghamshire	90.1%	81.9%	81.9%	87.4%
32 Lincolnshire	75.9%	74.6%	83.1%	76.4%
33 Leicestershire	78.9%	84.7%	83.5%	81.5%
34 Northamptonshire	89.0%	78.5%	87.7%	86.6%
35 Cambridgeshire	87.9%	80.1%	86.8%	85.7%
36 Norfolk	80.5%	83.2%	77.8%	80.6%
37 Suffolk	90.9%	84.8%	83.3%	85.7%
40 Bedfordshire	83.9%	80.7%	77.8%	81.5%
41 Hertfordshire	87.5%	78.3%	83.3%	83.9%
42 Essex	83.7%	76.7%	66.3%	80.3%
43 Thames Valley	84.4%	74.4%	80.0%	81.2%
44 Hampshire	87.4%	82.1%	79.9%	85.0%
45 Surrey	87.0%	80.0%	83.7%	83.8%
46 Kent	77.1%	78.1%	66.2%	76.2%
47 Sussex	81.2%	80.7%	86.8%	84.0%
48 City of London	48.4%	74.4%	56.3%	61.6%
50 Devon & Cornwall	91.4%	82.3%	89.1%	87.1%
52 Avon and Somerset	87.0%	82.3%	83.4%	84.4%
53 Gloucestershire	85.3%	83.1%	86.5%	84.9%
54 Wiltshire	77.4%	74.8%	82.9%	77.8%
55 Dorset	91.7%	81.2%	81.0%	86.9%
60 North Wales	84.8%	83.5%	80.6%	83.6%
61 Gwent	89.9%	87.2%	86.9%	88.5%
62 South Wales	83.9%	78.9%	81.5%	82.1%
63 Dyfed-Powys	86.0%	86.9%	83.1%	85.4%
TOTAL	82.5%	79.0%	82.7%	81.6%

Another issue of interest is whether the mean number of offences for each court date varies over the two data sources. To examine this issue, and to ensure comparability we looked at matched court dates only for our three major datasets Table 10.8 gives the results of this analysis. Before 1990, it appears that the PNC has the larger number of offences compared with the OI. However, from 1990 onwards, this pattern is reversed, and the OI contributes more

Table 10.8: Mean number of offences per court date in the matched data

		Before 1970	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999
PYO	O					2.75 [a]	3.71	3.45
	P					3.50 [a]	3.33	2.88
HDC	O	1.41	1.63	1.75	1.99	2.42	2.97	3.11 [b]
	P	1.71	1.88	2.07	2.06	2.55	2.88	2.63 [b]
Sentencing sample	O	1.33	1.62	1.87	2.06	2.29	2.65	2.55 [c]
	P	1.65	2.02	2.15	2.19	2.45	2.59	2.15 [c]
Overall	O	1.36	1.62	1.83	2.03	2.34	3.04	3.28
	P	1.67	1.97	2.12	2.14	2.49	2.88	2.75

[a] These figures are derived from very small numbers of offences. [b] For HDC the figures relate to 1995-2000. [c] For the sentencing sample, the figures relate to 1995-1997 only.

Key: O: OI data;
P: PNC data

offences per court date than the PNC. In interpreting this, it needs to be remembered that both datasets have been levelled to contain only standard list offences. The full PNC data, which includes non-standard list offences, will contain more offences by definition.

Finally, we examined common indigenous and non indigenous names, as we felt that common family names were more likely to be matched incorrectly, even if they have passed the matching process. The reasons for this have already been stated. The OI record might consist of composite individuals, with two John Smiths born on the same day being brought together into a single record. Alternatively, there might be twin brothers or sisters, who are carrying out crimes together and might be tried and sentenced on the same day from time to time. For non-indigenous family names such as Mohammed or Kim, unfamiliarity with the name might transpose family name with forename. Spanish and Maltese names have two family names and it can easily be a random choice which name become entered into the database.

To investigate this, we looked at the family names of SMITH, JONES, SINGH and ALI, and assessed the match rates at the court date level for the three large studies. The results are given below in Table 10.9, together with the number of individuals with each of these names in the link files.

Table 10.9: Matches at court date level by selected family names:

Study	Family name analysis				All family names
	Smith	Jones	Singh	Ali	
PYO	66.5% (132 cases)	64.7% (87 cases)	73.3% (4 cases)	31.8% (7 cases)	69.5%
HDC	73.6% (47 cases)	58.3% (29 cases)	87.5% (3 cases)	65.1% (8 cases)	74.5%
Sentencing sample	64.5% (58 cases)	63.4% (44 cases)	60.0% (3 cases)	54.8% (3 cases)	68.2%

There appeared to be strong evidence that the name Jones shows substantially poorer matching and Smith slightly poorer matching than other names. The family name of Ali also shows poorer matching performance, but there is no evidence that the matching rates for Singh are deflated.

11. Practical issues of hybridisation

11.1 What PNC records should be included on the OI?

The types of cases recorded on the PNC include convictions, cautions, reprimands, warnings and impending prosecutions. The offence information relating to these disposals can be summarised and matched or interleaved with OI disposal record information. Impending prosecutions may be valuable to researchers (though possibly not the provisional court dates) as alleged offence date information will allow arrest to be used as an additional outcome measure in evaluation studies. Thus all types of disposal, including court disposals known to the PNC but not known to the OI, would be included. Additionally, new records would be created for individuals not traced on the OI.

We also propose that summary details from all case types to be incorporated into the Offenders Index. Only those disposals relating to police forces in England and Wales (but including British Transport Police) should be covered; there is no intention of increasing the geographical coverage of the Offenders Index.

11.2 Specification of the hybridised record

We repeat the list of the types of information held by the PNC which was given in Section 1:

Type 1 -offender details (static)
Type 2 -some details of all proceedings
(whether convictions, cautions, reprimands, warnings or impending prosecutions and relevant date)
Type 3 -offence details
Type 4 –court disposal types and amounts
Type 5 –subsequent court appearance (eg varied on appeal) details
Type 6 –co-offenders

Pragmatic decisions need to be made as to which of this information to summarise and how to incorporate it into the Offenders Index. The proposal below has prioritised the data available for all disposals (type 1,2 and 3 records). The recording of offences in the PNC is more complete than the recording of court disposal types and amounts. For the recording of court disposal types and amounts it is proposed that where information has been matched and is available on both PNC and OI, then court data remains the only source i.e. not prioritising transferring court disposal (Type 4) information into the hybridised Offenders Index. What has been prioritised instead is the incorporation of co-offender details from the PNC into the hybridised Offenders Index. Further consideration is needed as to whether to give higher priority to incorporating PNC known court disposal details into a hybridised Offenders Index, though the model below does not.

In setting out how PNC data can be incorporated into the Offenders Index the following two strategic decisions have been applied:

1) Not altering the record type and column positions presently used in the Offenders Index to retrieve Offenders Index information using existing methodologies and code.
2) Summarising PNC records at the level 2 of the Offenders Index (disposal date). (The alternative of recording all PNC offences individually and the matching of all individual offences known to the Offenders Index would require much more developed offence coding).

The proposed structure of the hybridised Offenders Index - which retains the present three level structure- places the PNC disposal date details at Offenders Index level 2, to the right hand side of the existing data. Below, we set out the formats of the three data levels –followed by an example.

The structure at level 1 (record type 1) is proposed to remain as it presently is:

item	columns	Length	justification
Record type	1	1 character [ie '1']	
[space]	2		
OI Number	3-10	8 characters	right
[space]	11		
Surname	12-31	20 characters	left
[space]	32		
Initials	33-34	2 characters	left
[space]	35		
Date of birth	36-45	10 characters [DD/MM/YYYY]	
[space]	46		
Gender	47	1 character [1=male; 2=female]	
[space]	48		
Ethnicity1	49	1 character	
[space]	50		
CRO number2	51-59	9 characters	right
[space]	60		
Additional fields	61-268	209 characters	as input

Note that with the space for additional fields allowed in level 1 records, the width of a level 1 data file can be as wide as 268 characters. The structure proposed at level 2 extends to 288 characters. RDS needs to decide whether to trim or extend the fields proposed in the table below:

Item	columns	length	justification
Record type	1	1 character [ie '2']	
[space]	2		
Court data known disposal date	3-12	10 characters [DD/MM/YYYY]	
[space]	13		
Court data known Court code	14-17	up to 4 characters	right
[space]	18		
Court data known Curfew orders	19-21	up to 3 characters	right
[space]	22		
Court data known Date of previous disposal	23-32	10 characters [DD/MM/YYYY]	
[gap of 10 spaces]	33-42		
Court data known Age at disposal	43-44	2 characters	
[space]	45		
Court data known Number of previous disposals	46-48	up to 3 characters	right
[space]	49		
Court data known Number of subsequent disposals	50-52	up to 3 characters	right
[space]	53		
DATA SOURCE(S)	54	1 character [P=PNC,O=Offenders Index, B=Both]	
PNC Case Type	55	1 character [A=impending prosecution, C=Conviction, B=Caution, R=Reprimand, W=Warning]	
PNCID matched	56-68	up to 13 characters	left
[space]	69		
PNC Known disposal date	70-79	10 characters [DD/MM/YYYY]	
[space]	80		
PNC Known Court code	81-84	4 characters	right
[space]	85	[include leading zeroes above]	
PNC Known Curfew orders	86-88	up to 3 characters	right
[space]	89		
PNC Known Date of previous court disposal	90-99	10 characters [DD/MM/YYYY]	
[space]	100		
PNC Known Age at disposal	101-102	2 characters	
[space]	103		
PNC Known Number of previous disposals*	104-106	up to 3 characters	right
[space]	107		
PNC Known Number of subsequent disposals	108-110	up to 3 characters	right
[space]	111		
PNC Known total number of offences	112-114	up to 3 characters	right
[space]	115		
PNC Known total number of violence offences	116-118	up to 3 characters	right
[space]	119		
PNC Known total number of sexual offences	120-122	up to 3 characters	right
[space]	123		
PNC Known total number of burglary offences	124-126	up to 3 characters	right
[space]	127		
PNC Known total number of robbery offences	128-130	up to 3 characters	right
[space]	131		
PNC Known total number of theft offences	132-134	up to 3 characters	right
[space]	135		
PNC Known total number of forgery fraud offences	136-138	up to 3 characters	right
[space]	139		
PNC Known total number of criminal damage offences	140-142	up to 3 characters	right
[space]	143		

PNC Known total number of drugs offences	144-146	up to 3 characters	right
[space]	147		
PNC Known total number of other offences	148-150	up to 3 characters	right
[space]	151		
PNC Known total number of motoring offences	152-154	up to 3 characters	right
[space]	155		
PNC Known total number of uncategorised offences	156-158	up to 3 characters	right
[space]	159		
PNC Known number of offences committed on Bail*	160-162	up to 3 characters	right
[space]	163		
PNC Known earliest offence start date	164-173	10 characters [DD/MM/YYYY]	
[space]	174		
PNC Known latest offence start date	175-184	10 characters [DD/MM/YYYY]	
[space]	185		
PNC Known earliest offence end date	186-195	10 characters [DD/MM/YYYY]	
[Do we need end as well as start?] [space]	196		
PNC Known latest offence end date	197-206	10 characters [DD/MM/YYYY]	
[space]	207		
PNC Known Minimum Number of co-offenders at disposal date	208-210	up to 3 characters	right
[how far back do co-offender records go?][space]	211		
PNC Known Maximum Number of co-offenders at disposal date	212-214	up to 3 characters	right
[space]	215		
PNC Known Number of Different co-offenders at disposal date	216-218	up to 3 characters	right
[space]	219		
PNC Known median number of co-offenders at disposal date	220-222	up to 3 characters	right
[space]	223		
PNC Known Co-offender Lowest (earliest) PNCID *	224-236	up to 13 characters	right
[Not sure about some of these co-offender fields – would most frequent be better?][space]	237		
PNC Known Co-offender CRO for Lowest (earliest) PNCID	238-249	up to 12 characters	right
[space]	250		
PNC Known Co-offender Highest (latest) PNCID	251-263	up to 13 characters	right
[space]	264		
PNC Known Co-offender CRO for Highest (latest) PNCID	265-276	up to 12 characters	right
[space]	277		
PNC Known number of forces for disposal date	278	1 character	
[space]	279		
PNC Known Force code (minimum value for disposal date)**	280-281	2 characters	right
PNC Known Station code (attached to minimum force code)	282-283	2 characters	right
[space]	284		
PNC Known Force code (maximum value for disposal date)	285-286	2 characters	right
PNC Known Station code (attached to maximum force code)	287-288	2 characters	right

*an alternative would be to record the most frequent co-offender (though some disposal dates will have more than one most frequent co-offender)

** or most frequent police force (ditto, but less so)

Level 3 records would remain as they presently are:

item	columns	length	justification
Record type	1	1 character [ie '3']	
[space]	2		
Offence class code	3-5	3 characters	right
[space]	6		
Offence sub-class code	7-8	2 characters	right
[space]	9		
Police force code	10-11	up to 2 characters	right
[space]	12		
Proceedings type	13-14	up to 2 characters	right
[space]	15		
Plea	16	1 character	
[space]	17		
First disposal code	18-20	3 characters	right
[space]	21		
First disposal amount	22-25	4 characters	right
[space]	26		
First disposal units	27	1 character	
[space]	28		
Second disposal code	29-31	3 characters	right
[space]	32		
Second disposal amount	33-36	4 characters	right
[space]	37		
Second disposal units	38	1 character	
[space]	39		
Third disposal code	40-42	3 characters	right
[space]	43		
Third disposal amount	44-47	4 characters	right
[space]	48		
Third disposal units	49	1 character	
[space]	50		
Fourth disposal code	51-53	3 characters	right
[space]	54		
Fourth disposal amount	55-58	4 characters	right
[space]	59		
Fourth disposal units	60	1 character	
[space]	61		
Count of previous offences	62-64	3 characters	right
[space]	65		
Count of subsequent offences	66-68	3 characters	right

The example output file from the Offenders Index User Guide is reproduced on the following page. We have taken this example to illustrate how the OI would be altered after the above incorporation of PNC data into the OI. Due to the limitations of presenting 288 characters in the width of a page, the example output has been reproduced in three chunks; representing the left, middle and right hand side of the output. Only the left hand output is reproduced in this section; all three sections are shown in Appendix 4. All additional characters added to the Offenders Index in this example have been highlighted in bold.

CURRENT OI

```
1          HESTON               DJ 28/10/1957 1 0         0
2 0000000000          0000000000             000 000
3              0 000      0 000      0 000      0 000      0 000 000
1  6786013 CONNERY              AT 13/08/1956 1 1   8871393
2 01/11/1994 2932   0                     38 000 000
3  53 23 47  1 0 165  100 1 280    2 5 000    0 0 000    0 0 000 002
3 807  1 47  1 0 165  100 1 280    2 5 000    0 0 000    0 0 001 001
3 809  1 47  1 0 315  360 7 000    0 0 000    0 0 000    0 0 002 000
1  2097028 BROCOLI              DJ 20/06/1958 1 0   4791075
2 17/05/1974 6928   0                     15 000 013
3 918  1 47  2 1  01   10 £                                  000 025
3 918  1 47  2 1  01    5 £                                  001 024
2 17/10/1975  846   0 17/05/1974          17 001 012
3  30  0 47 50 3  21  730 D  03  160 £                      002 023
3  58 56 47 50 3  21  730 D                                 003 022
3 918  1 47 50 3  21  730 D                                 004 021
etc.
```

PROPOSED NEW OI

LEFT HAND SIDE OF DATASET (Characters 1-69)

```
1  9756011  HESTON              DJ 28/10/1957 1 0         0
2 0000000000          0000000000             000 000       PC1994/96
2 0000000000          0000000000             000 000       PC1994/96
3              0 000      0 000      0 000      0 000      0 000 000
1  6786013 CONNERY              AT 13/08/1956 1 1   8871393
2 01/11/1994 2932   0                     38 000 000       BC  1992/
3  53 23 47  1 0 165  100 1 280    2 5 000    0 0 000    0 0 000 002
3 807  1 47  1 0 165  100 1 280    2 5 000    0 0 000    0 0 001 001
3 809  1 47  1 0 315  360 7 000    0 0 000    0 0 000    0 0 002 000
2 0000000000          0000000000             000 000       PC  1992/
1  2097028 BROCOLI              DJ 20/06/1958 1 0   4791075
2 17/05/1974 6928   0                     15 000 013       C
3 918  1 47  2 1  01   10 £                                  000 025
3 918  1 47  2 1  01    5 £                                  001 024
2 17/10/1975  846   0 17/05/1974          17 001 012       C
3  30  0 47 50 3  21  730 D  03  160 £                      002 023
3  58 56 47 50 3  21  730 D                                 003 022
3 918  1 47 50 3  21  730 D                                 004 021
```

The first individual, HESTON, is someone who was not retrieved from the OI when researchers submitted his surname, initials, date of birth and gender. In this hypothetical example we are imagining there to have been a PNC record for DJ HESTON, which necessitates the creation of new OI level 2 records. The proposal is that a distinct range of OI numbers be set aside exclusively for OI records created from PNC information.

Part of the above specification proposes that the data source of the record is also recorded at level 2, with P indicating a disposal date (and details) known to the PNC alone, B indicating a date (including near matches) known to both data sources and O indicating a court disposal date known only from OI court data.

It is also proposed that additional court dates and other disposals from the PNC be added to the OI, including both cautions/warnings/reprimands and impending prosecutions and also court disposals not known to the OI.

It is important to note which of the data fields from the PNC contain information which are mandatory data fields on the PNC (see Appendix 1). It may be that some non-mandatory fields contain a substantial amount of missing data. As offence end date is not a mandatory field, we propose that the offence start date is placed in the offence end date position where no separate offence end date has been entered. This is because the majority of offences occur on a single day rather than over a period of time. This one amendment to the data as obtained from the PNC would probably have to be undertaken before the task of summarising all PNC offence records into disposal date summaries as set out above. As a consequence, it is recommended that offence start date becomes a mandatory field on the PNC, and is back record converted.

12. Exploration of the effect of hybridisation on a reconviction study

There are three main consequences of undertaking a reconviction study with a hybridised dataset as suggested in this report:

1) The number of known disposals (eg for Standard List Offences) increases;
2) The availability of date of offence information enables matched records to have pseudo-reconvictions excluded;
3) Any reconviction predictor tool (eg OGRS (Copas and Marshall, 1998)) having been calculated from a smaller dataset will under-predict reconviction rates known to the hybridised dataset,

In this section, we explore the effect of these changes on a typical reconviction study - the Strategic Alliance Reconviction Study- (Crosland and Rex, 2001). Table 12.1 shows some of the results. In this study, 1418 individuals received community sentences in the three targeted months of 1998 and were subsequently traced on the OI. The OI dataset contained OGRS scores for 1105 of these cases. Of these 1105 offenders, the number of individuals who had one or more subsequent court disposals within 2 years was found to be 544 i.e. 49.2%. Matching the 1105 individuals with those known to the PNC retrieved 1055 individuals with court disposals for standard list offences known to both data sources. Of these 1055 individuals, 567 were identified in the hybridised dataset to have had one or more court disposals within 2 years i.e. 53.7%. Thus, there appears to be an increase of 4.5% in the 2 year reconviction rate when the hybrid dataset is used. However, this is probably a slight under-estimation of the true figure as a few disposal details that should have been included in the hybrid dataset were excluded due to an error in the formatting of data sent by RDS. Whether this figure is typical for other reconviction studies can be reviewed more easily once PNC data is merged routinely into the OI. The degree to which different Police force areas differ should also be reviewed.

The meaning of reconviction rates obtained from the OI (especially for shorter reconviction periods) has always been jeopardised by not knowing how many of the reconvictions were for offences prior to the target conviction. A labour intensive calculation of pseudo-reconviction rates was undertaken in the work relating to Home Office Research Study 136 (Lloyd et al, 1994), and further modifications to the estimated pseudo-reconviction rate adjustments that need to be made in e.g. comparing community penalties to custody have been referred to in subsequent Home Office Publications e.g. Home Office Statistical Bulletin 19/1999. (Kershaw et al, 1999)

Table 12.1 The effect of pseudo-reconvictions on reconviction rates.

All Community Service Orders, Probation Orders and Combination Orders in four Probation Areas for 3 months in mid-1998 (N=1105)									
County	Data sample	N size	Individuals convicted within 2 years	% reconvicted within 2 years	individuals reconvicted within 2 years who had no recorded *offending* within 2 years	% of individuals reconvicted within 2 years who had no recorded offending within 2 years	pseudo reconviction adjustment to overall reconviction rate (for OI alone, see Note 1) (4)/(1)	% reconvicted within 2 years excluding pseudo-reconvictions ((2)-(4))/(1)	Net increase in known SLO reconvictions within 2 years (see Note 2)
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	
1	Hybrid -OI matched with PNC	251	114	45.42%	10	8.8%	3.98%	41.43%	6.12%
	OI alone	259	107	41.31%			6.00%	35.31%	
2	Hybrid -OI matched with PNC	265	152	57.36%	15	9.9%	5.66%	51.70%	4.44%
	OI alone	276	147	53.26%			6.00%	47.26%	
3	Hybrid -OI matched with PNC	290	132	45.52%	12	9.1%	4.14%	41.38%	7.19%
	OI alone	311	125	40.19%			6.00%	34.19%	
4	Hybrid -OI matched with PNC	249	169	67.87%	11	6.5%	4.42%	63.45%	5.75%
	OI alone	259	165	63.71%			6.00%	57.71%	
Total	Hybrid -OI matched with PNC	1055	567	53.70%	48	8.5%	4.55%	49.15%	5.92%
	OI alone	1105	544	49.23%			6.00%	43.23%	

Note 1 The standard estimate for community penalties is 6% reduction in overall reconviction rate -see Lloyd et al (1994); to adjust data relative to prison sentences the standard reduction is 4% see Statistical Bulletin 19/1999 (Kershaw et al, 1999)

Note 2 Of the standard offence list codes, 9 were not included in the dataset when they should have been -due to formatting errors in supplied data from RDS- so these figures are likely to be an underrecording of the SLOs known to the two data sources

Note 3 The measurement of 'reconviction' used for the OI was a count of court dates; it includes (somewhat unusually for reconviction research) breach proceedings.

There is some variability in the rates of pseudo-reconvictions found within the Strategic Alliance Research Study. For the 1055 offenders known to both PNC and OI, the rate of pseudo-reconvictions (to one decimal place) varied by Probation Area from 6.5% of reconvictions to 9.9% of reconvictions; the net effect on reconviction rates being from 4% to 5.7%. However, these pseudo-reconviction rates are not statistically different from one another (chi-squared=1.06 on 3 df; p=0.78). Overall, the 6% figure in current use seems robust for standard list offence pseudo-reconvictions.

Rates of reconviction set against predicted rates of reconviction are important to the evaluation of Criminal Justice work e.g. this measure is Key Performance Indicator One for all Probation Areas (under the National Probation Directorate). Having found some variation in pseudo-reconviction rates in different Police/Probation Areas, the assumption of uniform pseudo-reconviction rates should be revisited with larger datasets. Adding PNC data to the OI will enable an area by area re-evaluation of pseudo-reconviction rates over time and type of disposal. This allows the possibility of further development of prediction and evaluation tools such as OGRS.

13. Implications for reconviction research

13.1 Timeliness and completeness of reconviction studies

There are major benefits to researchers in having access to PNC data (separately or incorporated into the Offenders Index). The first is the reduced need to estimate pseudo-reconvictions. Additionally, there is more opportunity to consider alternative outcome measures such as arrests and other disposals such as cautions and warnings, which will make short follow-up times in evaluation studies more meaningful (Colledge, Collier and Brand,1999) Care, however, does need to be taken with fast follow-up studies to ensure that the dataset to be evaluated is likely to have the data on known disposals contained within it. A lag in the time for such data to appear on the PNC may cause problems:

> *"Following PITO advice, a two month 'buffer period' for reconvictions to reach the PNC has been adopted for interim evaluations. Using this buffer period provides more rapid results than are possible with the 6 to 9 month timelags associated with OI data. RDS recognise that more work needs to be done to assess the quality of PNC before it is used more widely,"(Howard and Kershaw,1999).*

However, from the PNC Performance figures published, it seems that the advice should clearly distinguish between the 'buffer period' for research depending on arrest data and the 'buffer period' for research looking at court disposals. The two month buffer for arrest data is supported by the October 1999 to October 2000 figures for entering 90% of arrest summons reports (the national average declining from 45 days to 36 days over this time period). Within this average there are of course variations by police force. In October 1999, for example, a four month buffer zone would have been necessary to capture more than 90% of arrest summons reports for Lincolnshire, Cheshire and Gwent police areas.

For court dates, the figures appear to be much worse. Between October 1999 and October 2000, the average number of days (from court date) to enter quickest 90% of court results went up from 196 days up to 206 days (though with one month of much better performance at 143 days). These figures are more difficult to interpret as the court date entered on the PNC for an impending prosecution is often earlier than the date on which the individual is actually sentenced. Thus, these figures can represent an overly pessimistic picture. However, this implies that a six month buffer period needs to be considered when considering reconvictions.

The Police National Computer at its best has court information on its system quickly enough for research into court disposals to take place from the PNC for some police areas after the two month 'buffer' period. Future researchers should be advised on this in the light of the relevant up to date PNC Performance figures.

The issue of data completeness also needs to be addressed. It must be borne in mind that though the OI and the PNC show evidence of having missing disposals and offences, there is much less evidence of offences being wrongly attributed to individuals who had not committed them. The exceptional cases, whilst individually of great concern, do not invalidate the analysis of offending data for large groups of individuals. Even if some details are missing, the disposals that are recorded overwhelmingly record real disposals for the individuals and make a study of known reoffending meaningful.

13.2 Auditing each data source from the other?

The incorporation of PNC data into the OI will allow auditing of each data source by using information from the other. The relative completeness of court data on the OI can be systematically and periodically checked against the PNC by looking for PNC court disposals not known to the OI. There is also the possibility of informing the Police of court data which is present on the OI and may be missing from their systems. If this information is fed back to the relevant Police Authorities in the form of research findings it would concord with the purpose and data protection registration of the Offenders Index. If it is thought that specific mismatches should be referred back to the PNC, there may well need to be legal advice sought before using the Offenders Index in this way. However, a caveat is needed – the OI uses far less stringent criteria for adding a conviction to an existing criminal history than does the PNC. While court data on an individual may well be missing on the PNC record, it may have been wrongly matched on the OI, or it might exist in a separate PNC record which cannot be matched to the first because of the lack of fingerprint information. The issue of audit therefore needs careful consideration.

13.3 Hybridisation models and updating of records.

If PNC information is to be placed on the Offenders Index, the decision as to how to carry out this hybridisation must be made. This needs to take into account the technical difficulty of the various methods and the value of the resulting data for various research projects.
There are two main methods which can be used:

a) A study by study match, with each new research dataset being processed and PNC data added, using procedures similar to those described in this report. This model will ensure that for a particular research study, all data is as up to date as possible.
b) A match of all existing OI data to all PNC records. However, in this case, updating of the hybridised dataset would still be needed, as new information would continue to be added to both the OI and PNC. This second option raises additional technical difficulties which have not been addressed in this report. In particular, the complexities of matching two large datasets together should not be underestimated. It is likely that such a matching procedure would only be carried out periodically, and research information from this source would be more out of date than with the previous model.

We describe in Section 14.2 a methodology for a more automatic method of matching court dates which could be used for either method.

13.4 Pseudo-reconvictions

The Review on Statistics of Efficacy of Sentencing comments :

> *"4.20 The adjustment for the effect of prior convictions on the reconviction rates of those sentenced to probation orders is of particular concern. It is a broad adjustment based on out of date small sample data and is larger than the apparent size of the difference one is trying to measure – i.e. the difference between reconviction rates between prison and probation order sentences. As a matter of priority every effort*

should be made to link the PNC and OI data in such a way as to allow the use of date of offence in place of date of conviction; this would also allow cautions to count as re-convictions. The current Lancaster/Cambridge University project (paragraph 4.11) should help to indicate how this might best be done." (Allnutt, 2001).

The 'pseudo-reconviction' terminology is actually inadequate. Some court disposals will have involved offending on either side of the date from which a reconviction study begins (the index date). These convictions will actually be 'part-pseudo, part-genuine reconviction' but this makes presentation somewhat difficult. As long as convictions are being counted and not offences, these cases can count as genuine reconvictions.

Some offences take longer to reach disposal than others; strengthening the case to move from reconviction to reoffending studies. However, the scope for undertaking studies of offences rather than just disposals within 2 years needs a methodological limitation to how long a researcher would wait to have a fair basis for including all offences within 2 years. Data needs to be produced periodically on the gaps between offence dates, arrest and disposal by offence type to gain information in this area.

13.5 Recalibrating reconviction predictors

The hybridised dataset will change known reconviction rates. This will be partly due to new court disposals entering the database from the PNC which are not known to the OI, and partly due to any change in definition of reconviction to include cautions, reprimands and warnings. The hybridised dataset will also make it possible to calculate 'known reoffending' rates. There is therefore a need to recalibrate reconviction predictor tools such as OGRS (Copas and Marshall, 1998). We see the need to update OGRS on a yearly basis to take account of changes in reconviction and reoffending rates over time. In 2002 a 2 year predictor based on 1999 disposals might be calculated–suggested name is OGRS1999a. A rolling programme could be established to review OGRS annually eg making OGRS2000a available in 2003. OGRS1999a might, for example, be a predictor for Standard List Offence disposals (excluding breaches). Other prediction models can also be developed based on different combinations of :

- Disposals (eg. court/other)
- Follow-up period (eg 6,12,18 months)
- Offence types (eg standard list/non-standard list).

13.6 The need for new research guidelines.

The above changes and developments will mean that researchers and other users of the hybridised dataset will need advice and guidelines on how to use the dataset, as well as a discussion of some of the issues outlined in this report. We recommend that RDS develops best practice guidelines for evaluation studies and other conviction research based on the enhanced dataset, describing the use of the new information. As well as describing the use of such fields as data of offence. the limitations of each of the constituent parts of the dataset should be recognised as outlined in Sections 4 and 5.

In addition, RDS can improve its service to researchers by ensuring that every future data file supplied (OI, PNC or hybridised) is provided with an electronic version of the data structure and all coding used (with the latest date of change of the data structure and any of the coding recorded). Data positions of researchers additional fields should be provided.

14. Developing an automatic matching system for court dates

The procedure for matching described in the early part of the report was devised for the purpose of comparing two levelled datasets to assess robustness and data completeness. The software developed relies on a substantial amount of user intervention to bring court date records together where the dates may disagree. For operational use, it would be useful to devise a more automatic system. This section describes a statistical analysis of the outcome of the matching process which can be used to construct a matching score for any two court dates and their associated offences. The method is based on work by Newcombe (1988) and Wain, Francis and Stott (1993). We first describe the statistical analysis and the construction of the score. We then describe the process of validation, and propose an operational system which could be used in practice to match OI and PNC datasets at the court date level.

14.1 The statistical analysis

We take as a starting point the set of all matched court dates for the Home Detention Curfew sample. Table 10.4 identifies that there are 26864 matched court dates in this dataset, with associated information on police codes, court codes and offence summaries from both the PNC and OI as described in Section 8. We carry out the following operations on a slightly smaller dataset of 26862 court records:

a) We divide this dataset into two approximately equal sized subsamples. The first subsample remains unchanged and represents a set of 'true matched' court dates. The second half of the dataset is modified to contain a set of court dates which are 'true mismatches'. We do this by randomly sorting the PNC dates and associated information and linking record for record with the unsorted OI dates and information. The purpose of this randomisation is to provide a sample of mismatched records to allow the statistical analysis to distinguish between mismatches and matches. The occasional match might occur by chance in the randomisation subsample but we ignore this possibility as it will have little effect on the statistical analysis. (The phrases 'true match' and 'true mismatch' in fact represent the decisions reached by the manual matching process rather than any absolute notion of truth).

b) Each subsample is then again divided into two - 60% of each subsample is defined to be a training subset and is used in the statistical analysis, and the remaining 40% of each subsample is defined to be a validation subset used to assess model performance.

This is represented diagrammatically in Figure 14.1. Subsets A and C are the training subsets, and subsets B and D are the validation subsets.

Figure 14.1 Division of the HDC matched court date records into four subsets.

MATCHES				***MISMATCHES***		
subsample 1	*OI data*	*PNC data matched to OI*		*subsample 2*	*OI data*	*randomised PNC data*
subset A 8093 records			*training sample 60%*	*subset C 8029 records*		
subset B 5338 records			*validation sample 40%*	*subset D 5402 records*		

Logistic regression was then used to model the probability of the OI court record matching the PNC court record, using subsets A and C. The response was defined to be an indicator variable MATCH taking the value 1 if the record was an OI-PNC match (subset A) and zero if the record was an OI-PNC mismatch (subset C). A comprehensive set of explanatory variables was constructed and are listed in Table 14.1.

Table 14.1 Explanatory variables used in the logistic regression

Name of variable	Type	Definition
A_NOFFV	indicator	number of violent offences between PNC and OI disagree=0; agree=1
D_NOFFV	continuous	absolute difference in number of violent offences between PNC and OI
A_NOFFS	indicator	number of sexual offences between PNC and OI disagree=0; agree=1
D_NOFFS	continuous	absolute difference in number of sexual offences between PNC and OI
A_NOFFB	indicator	number of burglary offences between PNC and OI disagree=0; agree=1
D_NOFFB	continuous	absolute difference in number of burglary offences between PNC and OI
A_NOFFR	indicator	number of robbery offences between PNC and OI disagree=0; agree=1
D_NOFFR	continuous	absolute difference in number of robbery offences between PNC and OI
A_NOFFT	indicator	number of theft offences between PNC and OI disagree=0; agree=1
D_NOFFT	continuous	absolute difference in number of theft offences between PNC and OI
A_NOFFF	indicator	number of fraud/forgery offences between PNC and OI disagree=0; agree=1
D_NOFFF	continuous	absolute difference in number of fraud/forgery offences between PNC and OI
A_NOFFC	indicator	number of criminal damage offences between PNC and OI disagree=0; agree=1
D_NOFFC	continuous	absolute difference in no. of criminal damage offences between PNC and OI
A_NOFFD	indicator	number of drugs offences between PNC and OI disagree=0; agree=1
D_NOFFD	continuous	absolute difference in number of drugs offences between PNC and OI
A_NOFFM	indicator	number of motoring offences between PNC and OI disagree=0; agree=1
D_NOFFM	continuous	absolute difference in number of motoring offences between PNC and OI
A_NOFFO	indicator	number of other offences between PNC and OI. disagree=0; agree=1
D_NOFFO	continuous	absolute difference in number of other offences between PNC and OI
A_DAYOC	indicator	day of court date on PNC and OI. disagree=0; agree=1
DIFFDAYC	continuous	absolute difference in day of court date on PNC and OI
A_MONCON	indicator	month of court date on PNC and OI. disagree=0; agree=1
DIFFMONC	continuous	absolute difference in month of court date on PNC and OI
A_YEACON	indicator	year of court date on PNC and OI. disagree=0; agree=1
DIFFYEAC	continuous	absolute difference in year of court date on PNC and OI
A_CRT_MS	indicator	smallest court code on OI and PNC agrees or PNC code missing (1=yes 0=no)
A_CRTS_3	indicator	smallest court code on OI and PNC agrees in last three digits (1=yes 0=no)
A_CRTS	indicator	smallest court code on OI and PNC agrees totally (1=yes 0=no)
A_POL_MS	indicator	smallest police code on OI and PNC agree or OI code missing (1=yes 0=no)

The court code and police code variables need some explanation. It is possible when summarising offending histories by court date, for more than one court code to be present in the database. The matching software summarised court code information into two variables – the smallest court code and the largest court code within a court date. A similar

process was undertaken for police code. Mostly, the smallest court code and the largest court code have the same value, but occasionally they differ. Magistrates court codes have four digits with the first digit representing whether the court is a youth court or an adult court. Court codes often agree on the last three digits but disagree in the first digit. Additionally, court codes are often missing on the PNC and are coded 9998 if this is so. Occasionally OI police codes are also missing.

A forward stepwise logistic regression procedure was adopted to find the best predictive model for the probability of a match between a PNC record and an OI record. At each stage, the explanatory variable giving the largest decrease in minus twice the log-likelihood when added to the model was included, providing the change was significant (that is, greater that the criterion value of 3.84 (the 95% percentile of the chi-squared distribution on one degree of freedom). The procedure added thirteen explanatory variables to the final model.

The estimates are shown in Table 14.2. It can be seen that all remaining variables in the model are significant according to the Wald test ($p<0.02$ for any variable).

Table 14.2 Parameter estimates from the final model of the logistic regression

	parameter estimate	standard error of estimate	Wald test p-value.
Constant	-14.004	1.519	.000
A_NOFFV	1.431	.409	.000
A_NOFFB	1.217	.381	.001
A_NOFFR	2.314	.953	.015
A_NOFFT	1.749	.346	.000
A_NOFFF	2.287	.677	.001
A_NOFFC	1.448	.565	.010
D_NOFFO	-.530	.146	.000
A_DAYOC	3.878	.440	.000
A_MONCON	3.445	.420	.000
DIFFYEAC	-2.290	.268	.000
A_CRT_MS	1.178	.379	.002
A_CRTS_3	2.486	.538	.000
A_POL_MS	5.162	.439	.000

The parameter estimates can be used to build a classification score defined by calculating the predicted logit score from the estimates. This is constructed as follows:

a) start from the constant of –14.004
b) if there is agreement in the number of violent offences add 1.431 otherwise add 0
c) if there is agreement in the number of burglary offences add 1.217 otherwise add 0
d) if there is agreement in the number of robbery offences add 2.314 otherwise add 0
e) if there is agreement in the number of theft offences add 1.749 otherwise add 0
f) if there is agreement in the number of fraud/forgery offences add 2.287 otherwise add 0
g) if there is agreement in the number of criminal damage offences add 1.448 otherwise add 0
h) calculate the absolute difference in the number of other offences, multiply by 0.530 and subtract the result
i) if there is agreement in the day of the court disposal add 3.878 otherwise add 0
j) if there is agreement in the month of the court disposal add 3.445 otherwise add 0
k) calculate the absolute difference in the year of court disposal, multiply by 2.290 and subtract the result

l) if the court codes agree in the last three places add 2.486
or if the PNC court code is missing add 1.178
or if the court codes agree totally add 3.664

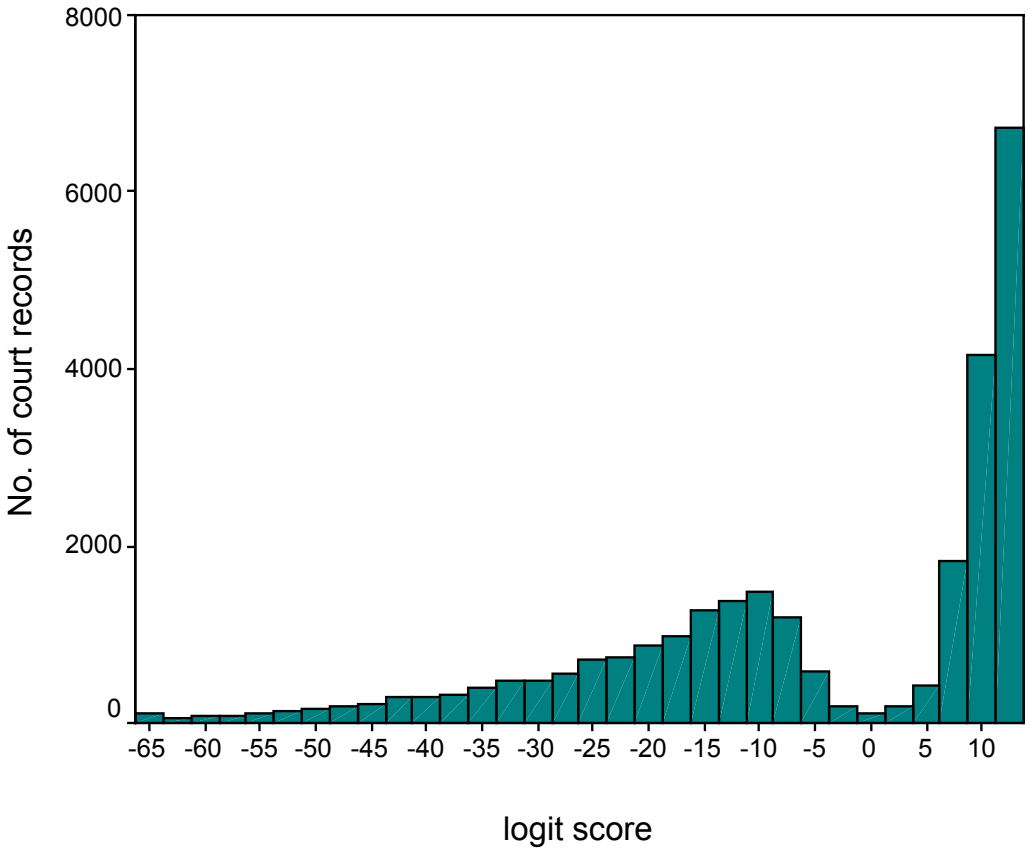
m) if the police codes agree or the OI code is missing add 5.162

If the resulting logit score is greater than zero, then we accept the two court records as agreeing, if it is less than zero, then we reject the two court records as a match. We choose a logit score of zero as the cut point as this represents an equal chance of the OI and PNC records being a true match and being a false match. This is explained by the fact that logit scores can be transformed into probabilities using the mathematical equation

$$\text{Probability of an OI and a PNC records being a 'true' match} = \frac{\exp(\text{logit score})}{1+\exp(\text{logit score})}$$

Figure 14.1 shows the logit scores for all 26862 court records.

Figure 14.1 Estimated logit scores for HDC study – all court-level records.



We can see that the logit score distribution has two peaks – one a high positive value around 15 and the other a high negative value around –10. At the cut point of zero, there are few scores – the logit score divides the data well into two groups.

To assess how well this score performs, we look at the validation sample (subsets B and D). If the score performs perfectly, we would expect all court records in subset B to get a positive score, and all court records in subset D to get a negative score. Table 14.3 shows the actual result for this validation sample.

Table 14.3 Applying the logit score to the HDC validation sample

	negative logit score (predicted mismatch)	positive logit score (predicted match)	Total
Subset D mismatched records	5325	13	5338
subset B matched records	25	5377	5402
Total	5350	5390	10740

The logit score performs excellently in reproducing the results of the manual matching process. 10702 out of 10740 records are classified correctly – a success rate of 99.65%. The misclassification rate is 0.35% - only 38 records are misclassified compared with the results of the manual matching process.

We have indicated above that the logit scores can be transformed into probabilities. If we carry out this procedure, we obtain the results given in Table 14.4:

Table 14.4 Estimated probabilities of a match for true matches and mismatches HDC validation sample

	'true' matches	'true' mismatches	Total
$p<0.001$ or $p>0.999$	4910	5084	9994
$p<0.1$ or $p>0.9$	5353	5303	10656
$0.1 \leq p \leq 0.9$	49 (0.9%)	35 (0.7%)	84 (0.8%)
Total	5402	5338	10740

The vast majority of PNC-OI pairs have probabilities less than 0.1 or greater than 0.9 – these cases can be regarded as assigned with high probability to 'mismatch' or 'match'. If we treat those pairs where the probability lies between 0.1 and 0.9 as those records where there may be some doubt in allocating a record pair to the matched or unmatched group, these pairs still constitute less than one in a hundred cases. These cases however may need looking at manually.

Even though we have measured success on the HDC validation sample and not on the training sample, it might be argued that we are still using the same study (the Home Detention Study) to assess success. How does the score perform on another of the studies? We therefore applied the score to the Sentencing sample records. From Table 10.4, we note that there were 27937 court dates matched by the manual process. How many of these have positive logit scores using the above formula? Table 14.4 below shows the result.

Table 14.4 Applying the logit score to the sentencing sample matched court dates

	negative logit score (predicted mismatch)	positive logit score (predicted match)	Total
matched records	112	27825	27937
%	0.40	99.60	100.0

The result is again impressive, with nearly the same success rate as for the HDC validation sample.

14.2 An outline of an operational system.

With the above work, an outline of a proposed classification system can be proposed. We describe the system as though a set of names have been presented to RDS as part of a research study (Figure 2.1)

1. For every offender under study, trace the offender on the PNC and the OI and obtain the criminal histories.
2. Exclude non England and Wales convictions from the PNC if this is desired.
3. For every distinct court date on the PNC and on the OI, calculate the number of standard list offences in each of the ten Criminal Statistics categories. Also calculate the smallest court code and police code used for that court date.
4. For **each** court date on the Offender Index, calculate the logit score to **every** court date with a court conviction on the PNC. Accept the PNC court date record with the highest logit score as a match if the score is above zero, otherwise treat the court date as unmatched. Flag the PNC record which has been matched as taken, and do not consider it for subsequent matching for this offender.
5. Merge in all unmatched PNC records, including all unmatched court convictions, and all records relating to non-court disposals.
6. (Optionally) manually examine all offenders with low match rates (less than 50%) to see if there is evidence of split records on the OI which need demerging.

It remains to be seen how this procedure performs in practice. If more time were available, this would be the next stage of the research project. In particular, Stage 4 of the above procedure may be too time consuming and other less exact methods may be acceptable.

We recommend that RDS pilots the above procedure and evaluates it on new and existing datasets. This will again need purpose written software to be constructed.

A final problem when matching records is that it is easy to obtain a hybridised record with contradictory information in the PNC half and the OI half of the record. Which information should be taken? This needs further investigation – our provisional thoughts are that:

- Individual level fields such as gender and date of birth should be reconciled where possible. It is however likely that the PNC contains more accurate data.
- Differences in court dates matter less if date of offence is to be used as the outcome measure in reconviction studies. It is likely that the OI as the primary source contains more accurate information.

- Contradictory information is likely to exist in the number and type of offences. Both PNC and OI data are likely to be reliable – the advice here is to choose the OI information if standard list offences are required and PNC if all offences are needed.
- Information on court disposal types and amounts is best obtained from the OI.

Acknowledgements

The authors are grateful to Philip Howard and Julian Prime, who commissioned this research and who provided much helpful guidance and encouragement throughout the research. At Lancaster, Juliet Harman dealt with the preparation, recoding and analysis of the data studies and we are particularly grateful for her skill, accuracy and dedication. We also thank Jayn Pearson, who has worked tirelessly in carrying out the majority of the matching work.

References

Allnutt, D (2001) Review of Statistics on Efficacy of Sentencing. Office of National Statistics, London
http://www.statistics.gov.uk/themes/yourviews/downloads/postconsdrft.pdf

Colledge, M., Collier, P. and Brand, S. (1999). Programmes for Offenders: Guidance for Evaluators. Crime Reduction Programme, Guidance Note 2, London: Home Office.

Copas, J. and Marshall, P. (1998). *'The Offender Group Reconviction Scale: the statistical reconviction score for use by Probation Officers*.' Journal of the Royal Statistical Society, Series C, 47, 159-171.

Crosland, P. (1999) *Report on Offenders' Index and Police National Computer Data for the Strategic Alliance Steering Group* (unpublished)

Crosland, P. and Rex, S. (2001) *The Strategic Alliance Probation Areas Reconviction Study*, Herts, Northants, Beds and Cambs Probation Services.

Dodgson. K et al. (2001) . *Electronic monitoring of released prisoners: an evaluation of the Home Detention Curfew scheme.* Home Office Research Study No 222, London: HMSO.

Friendship, C.et al (2000) *Reconviction: a critique and comparison of two main data sources in England and Wales* (submitted to Legal and Criminological Psychology, March 2000)

Home Office (2000) Criminal Statistics, England and Wales 1999. London:HMSO

Howard, P. and Kershaw, C. (2000) *Using Criminal Career Data in Evaluation* in Mair, R. and Tarling (2000) The British Criminology Conference: Selected Proceedings. Volume 3. Papers from the British Society of Criminology Conference, Liverpool, July 1999.

Kershaw, C. et al (1999) *Reconvictions of offenders sentenced or discharged from prison in 1995, England and Wales.* Home Office Statistical Bulletin 19/1999

Lloyd, C., Mair, G. and Hough, M. (1994) *Explaining Reconviction Rates: A Critical Analysis.* Home Office Research Study No 136, London: HMSO.

Newcombe, H.B. (1988) *Handbook of Record Linkage*. Oxford University Press, Oxford

Police Information Technology Organisation (2000) *Phoenix System Data Definitions Version 3.23.* Police National Computer Directorate.

Povey, K (2000) *On The Record –Thematic Inspection report. Police crime recording, the Police National Computer, Phoenix Intelligence System Data Quality* Her Majesty's Inspectorate of Constabulary, London.

Prime, J. et al. (2001) *Criminal careers of those born between 1953 and 1978* Statistical Bulletin 4/01 *London:* Home Office.

Rose, G, (2000) *The Criminal Careers of Serious Traffic Offenders* Home Office Research Study No 206, London: HMSO.

Russell, J. (1998) *Phoenix Data Quality*. Special Interest Series: Paper 11. London: Police Research Group.

Soothill, K., Francis, B., Ackerley, E. and Sanderson, B. (2000). Sex offenders: specialists, generalists or both?: a 32-year criminological study. *British Journal of Criminology*, **40**, 56--67.

Tendler, S (2001) *Records chaos may put police chiefs in court* The Times, 14th April, p15.

Wain, R, Francis, B and Stott, D. (1993) Software for routine record linkage in public health In: *Healthcare Computing 1994* (Ed. Richards, B) pp 681-688. ISBN 948198 17 6

Appendix 1-Mandatory PNC data fields supplied to RDS under existing RDS/PITO protocol:

Record Type 1 –Individual Details

PNCID

PNC Filename

Sex

Date of Birth

Ethnic Appearance

BRC Status [Back Record Conversion] *

Record Type 2-Proceedings (Impending prosecution; caution; reprimand; warning; caution)

[None]

Record Type 3 –Offence

Offence Code

Offence start date

Force Bringing Charges

Record Type 4-Disposal (only present after sentencing)

Disposal Type

Record Type 5- Subsequent Appearance

Date of Subsequent Appearance

Court code for Subsequent Appearance

Record Type 6- Co-offender

PNCID of Co-offender

Name of Co-offender

*Back Record Conversion Status codes are as follows:

```
O     CREATED POST-MIGRATION
N     NO CONVICTIONS ON PNC
C     PNC2 CONVICTIONS LEVEL ONLY
E     ENHANCED LEVEL
F     FULL POLICING DATA
```

Appendix 2 –How Magistrate Court information gets to the Court Proceedings Database

The 23 Police Forces presently responsible for providing Magistrates Court records to the Court Proceedings Database	The 20 Police Force Areas where the Magistrates Court data system directly supplies the Court Proceedings Database
Bedfordshire	Avon and Somerset
Cambridgeshire	Cheshire
City of London	Devon and Cornwall
Cleveland	Dorset
Cumbria	Essex
Derbyshire	Hampshire
Durham	Humberside
Gloucestershire	Lancashire
Greater Manchester	Leicestershire
Hertfordshire	Lincolnshire
Kent	Northamptonshire
Merseyside	Northumbria
Metropolitan Police	Nottinghamshire
Norfolk	Suffolk
North Yorkshire	Sussex
South Yorkshire	West Yorkshire
Staffordshire	Wiltshire
Surrey	Dyfed Powys
Thames Valley	Gwent
Warwickshire	South Wales
West Mercia	
West Midlands	
North Wales	

Appendix 3 Preparing the data for comparative matching

Matching at individual level

1. Identify and delete duplicated individuals in the link file.
2. Delete duplicated records from the raw OI and PNC data files.
3. Match individuals by merging the link file with the PNC file aggregated on PNCID and with the OI file using link file ID.

Matching sentencing occasions and number of offences

4. Identifying those data in the both files which are expected to be known to the OI using the following match criteria for each offence:

    a) match criterion 1 – conviction

    b) match criterion 2 – standard list offence

    c) match criterion 3 – caution/conviction date not too recent for match

    d) match criterion 4 – sentenced within same jurisdiction as covered by comparison data

5. Create a 'level playing field' by excluding these records from the match of both files.

The number of offence level records excluded are shown below; from these figures some indication of how many additional records might be in a hybrid dataset may be estimated. For example, note the 22% of records excluded from the PNC due to not being recognised by the coding as Standard List Offences –as opposed to just 3% of the Offenders Index offence records.

Sentencing sample

	PNC	OI
At offences record level	211,835 records	148,288records
met match criterion 1 (conviction)	94.8%	all
met match criterion 2 (S.L.O.)	80.6%	94.9%
met match criterion 3 (not too recent for match)	76.6%	99.9%
met match criterion 4 (valid Police Force for matching process)	98.2%	100%
met all 4 match criteria	125,195 records (59.1%)	140,505 records (94.8%)
At court date level	96,361 records	65,775 records
met all 4 match criteria	56,440 records (58.6 %)	60,820 records (92.5%)

PYO files

	PNC	OI
At offences record level	299,593 records	245,182 records
met match criterion 1 (conviction)	91.0 %	all
met match criterion 2 (S.L.O.)	82.9 %	94.2%
met match criterion 3 (not too recent for match)	96.9 %	99.8%
met match criterion 4 (valid Police Force for matching process)	99.0 %	100%
met all 4 match criteria	222,707 records (74.3%)	230,948 records (94.2%)
At court date level	123,122 records	79,949 records
met all 4 match criteria	81,023 records (65.8%)	70,563 records (88.3%)

Strategic Alliance

	PNC (Nov. 2000)	PNC (June 2001)	OI
At offences record level	28,166 records	28,861 records	21,869 records
met match criterion 1 (conviction)	92.9%	93.5%	all
met match criterion 2 (S.L.O.)	77.4%	77.8%	94.3%
met match criterion 3 (not too recent for match)	97.4%	94.1%	99.2%
met match criterion 4 (valid Police Force for matching process)	98.6%	98.6%	100%
met all 4 match criteria	19,832 records (70.4%)	19,913 records (69.0%)	20,527 records (93.9%)
At court date level	13,041 records	13,285records	9,257 records
met all 4 match criteria	8,735 records (67.0%)	8,743 records (65.8%)	8,349 records (90.2%)

HDC

	PNC	OI
At offences record level	310,996 records	232,818 records
met match criterion 1 (conviction)	95.6%	all
met match criterion 2 (S.L.O.)	78.2%	95.8%
met match criterion 3 (not too recent for match)	97.2%	99.7%
met match criterion 4 (valid Police Force for matching process)	98.4%	100%
met all 4 match criteria	226,357 records (72.8%)	222,974 records (95.8%)
At court date level	132113 records	89,932records
met all 4 match criteria	91944 records (69.6 %)	83,924records (93.3%)

Teesside Pathfinder

	PNC	OI
At offences record level	23919 records	18195 records
met match criterion 1 (conviction)	94.8%	all
met match criterion 2 (S.L.O.)	81.0%	96.7%
met match criterion 3 (not too recent for match)	95.7%	99.8%
met match criterion 4 (valid Police Force for matching process)	99.7%	100%
met all 4 match criteria	18236 records (76.2%)	17592 records (96.7%)
At court date level	9899 records	7016 records
met all 4 match criteria	7237 records (73.1%)	6668 records (95.0%)

Devon Pathfinder

	PNC	OI
At offences record level	6775 records	4647 records
met match criterion 1 (conviction)	96.4 %	all
met match criterion 2 (S.L.O.)	79.0 %	96.0%
met match criterion 3 (not too recent for match)	97.8 %	99.5%
met match criterion 4 (valid Police Force for matching process)	98.1 %	100%
met all 4 match criteria	4998 records (73.8%)	4460 records (96.0%)
At court date level	2610 records	1829 records
met all 4 match criteria	1855 records (71.1%)	1717 records (93.9%)

Hereford + Worcester Pathfinder

	PNC	OI
At offences record level	5952 records	3908 records
met match criterion 1 (conviction)	90.6 %	all
met match criterion 2 (S.L.O.)	86.7 %	93.6%
met match criterion 3 (not too recent for match)	97.0 %	97.4%
met match criterion 4 (valid Police Force for matching process)	99.2 %	100%
met all 4 match criteria	4622 records (77.7%)	3656 records (93.6%)
At court date level	2638 records	1614 records
met all 4 match criteria	1834 records (69.5%)	1441 records (89.3%)

Appendix 4: Output file for Model of proposed incorporation of PNC records into Offenders Index

LEFT HAND SIDE OF DATASET (Characters 1-69)

```
1  9756011 HESTON               DJ 28/10/1957 1 0         0
2 0000000000          0000000000              000 000       PC1994/96
2 0000000000          0000000000              000 000       PC1994/96
3              0 000      0 000      0 000      0 000      0 000 000
1  6786013 CONNERY              AT 13/08/1956 1 1   8871393
2 01/11/1994 2932   0                      38 000 000       BC  1992/
3  53 23 47  1 0 165  100 1 280    2 5 000    0 0 000    0 0 000 002
3 807  1 47  1 0 165  100 1 280    2 5 000    0 0 000    0 0 001 001
3 809  1 47  1 0 315  360 7 000    0 0 000    0 0 000    0 0 002 000
2 0000000000          0000000000              000 000       PC  1992/
1  2097028 BROCOLI              DJ 20/06/1958 1 0   4791075
2 17/05/1974 6928   0                      15 000 013       C
3 918  1 47  2 1  01   10 £                                   000 025
3 918  1 47  2 1  01    5 £                                   001 024
2 17/10/1975  846   0 17/05/1974          17 001 012       C
3  30  0 47 50 3  21  730 D  03  160 £                       002 023
3  58 56 47 50 3  21  730 D                                  003 022
3 918  1 47 50 3  21  730 D                                  004 021
```

MIDDLE OF DATASET (Characters 54-173)

```
PC1994/9601223M 08/05/1994 2003     0000000000 36 000 001 006 000 000 000 000 000 000 001 000 001 003 001 000 30/03/1994
PC1994/9601223M 10/06/1996 2003     08/05/1994 38 001 000 004 000 000 000 000 002 002 000 000 000 000 000 000 08/09/1995
0 000 000

BC  1992/56401P 02/11/1994 2932     0000000000 38 000 001 003 000 000 000 000 000 001 000 000 000 002 000 000 28/09/1993
0 000 002
0 001 001
0 002 000

PC  1992/56401P 05/12/1995 2932     02/11/1994 39 001 000 002 000 000 000 000 000 000 000 000 000 000 002 000 000 22/08/1995

C
000 025
001 024
002 023
003 022
004 021

C
005 020
006 019
007 018
008 017
009 016
010 015
```

RIGHT HAND SIDE OF DATASET (Characters 164-288)

```
30/02/1994 05/03/1994 30/02/1994 07/03/1994 000 002 002 000  1988/568947K   568947/88X 1994/6785222L    25781/94V 1 04C1 04C1
08/09/1995 08/09/1995 08/09/1995 08/09/1995 000 000 000 000                                                 1 04C1 04C1


28/04/1994 28/04/1994 28/04/1994 28/04/1994                                                                 1 47B2 47B2



22/08/1995 22/08/1995 22/08/1995 22/08/1995                                                                 1 47B2 47B2
```

Appendix 5 The 108 acts with ACPO or CCCJS offence codes treated as missing on conversion

Agricultural Marketing Act 1958
Agriculture Act 1967
Agriculture Act 1970
Air Force Act 1955
Air Navigation (No2) Order 1995
Alcoholic Liquor Duties Act 1979
Animals (Scientific Procedures Act) 1986
Army Act 1955
Aviation and Maritime Security Act 1990
British Telecommunications Act 1981
Broadcasting Act 1990
Census Act 1920
Common Law
Companies Act 1985
Company Directors Disqualification Act 1986
Company Securities (Insider Dealing) Act 1985
Control of Pollution (Silage, Slurry and Agricultural fuel Oil) Regulations 1991
County Courts Act 1984
Cremation Act 1902
Crime (Sentences) Act 1997
Criminal Justice Act 1961
Criminal Justice Act 1987
Criminal Justice Act 1988
Customs and Excise Duties (General Reliefs) Act 1979
Data Protection Act 1984
Debtors Act 1869
Estate Agents Act 1979
Explosive Substances Act 1883
Explosives Act 1875
Factories Act 1961
Fair Trading Act 1973
Finance Act 1989
Financial Services Act 1986
Fire Precautions Act 1971
Fishing Boats (Marking and Documentation)(Enforcement) Order 1993
Food and Environment Protection Act 1985
Food Safety Act 1990
Forgery Act 1861
Gas Act 1965
Genocide Act 1969
Gun Barrel Proof Act 1868
Housing Act 1964
Incitement to Disaffection Act 1934
Industrial Training Act 1982
Insolvency Rules 1986
Interception of Communications Act 1985
Internationally Protected Persons Act 1978
Iron and Steel Act 1982
Marine etc Broadcasting Offences Act 1967
Marriage Act 1949
Medicines (Advertising of Medicinal Products) (No 2) Regulations 1975
Medicines (Advertising of Medicinal Products) Regulations 1975
Medicines (Advertising to Medical and Dental Practitioners) Regulations 1978
Medicines (Child Safety) Regulations 1975
Medicines (Contact Lens Fluids and Other Substances) (Advertising and Miscellaneous Amendments) Regulations 1979
Medicines (Contact Lens Fluids and Other Substances) (Labelling) Regulations 1979
Medicines (Fluted Bottles) Regulations 1978
Medicines (Labelling and Advertising to the Public) Regulations 1978
Medicines (Labelling) Regulations 1976
Medicines (Leaflets) Regulations 1977
Medicines Act 1968
Medicines for Human Use (Marketing Authorisations etc) Regulations 1994
Merchant Shipping (Safety Convention) Act 1949
Merchant Shipping Act 1894
Mock Auctions Act 1961
Nuclear Material (Offences) Act 1983
Official Secrets Act 1911
Official Secrets Act 1989
Piracy Act 1698
Piracy Act 1721
Piracy Act 1837
Population (Statistics) Act 1938
Prevention of Corruption Act 1906
Prevention of Oil Pollution Act 1971
Prison Act 1952
Prison Security Act 1992
Prohibition of Female Circumcision Act 1985
Public Bodies Corrupt Practices Act 1889
Public Health Act 1936
Rabies (Importation of Dogs, Cats and other Mammals Act) 1974
Radioactive Substances Act 1948
Radioactive Substances Act 1960
Registered Designs Act 1949
Registered Homes Act 1984
Representation of the People Act 1983
Reservoirs Act 1975
Rivers (Prevention of Pollution) Act 1961
Salmon Act 1986
Sea Fish (Conservation) Act 1967
Sea Fish Industry Act 1970
Sexual offences(Conspiracy and Incitement) Act 1996
Sheriffs Act 1887
Solicitors Act 1974
Statutory Declaration Act 1835
Taking of Hostages Act 1982
Taxes Management Act 1970
Telegraph Act 1868
Theatres Act 1968
Trade Descriptions Act 1968
Trade Marks Act 1938
Trading Representations (Disabled Persons) Act 1958
Trading Stamps Act 1964
Unlawful Drilling Act 1819
Venereal Diseases Act 1917
Veterinary Surgeons Act 1966
Water Industry Act 1991
Water Resources Act 1991
Water Supply (Water Quality) Regulations 1989

Appendix 6 The 42 acts with ACPO or CCCJS offence codes represented by a hyphen on conversion

Animal Health Act 1981
Badgers Act 1973
Banking Act 1979
Bankruptcy Act 1914
Cable and Broadcasting Act 1984
Child Benefit Act 1975
Child Care Act 1980
Coinage Offences Act 1936
Companies Act 1948
Companies Act 1976
Companies Act 1980
Companies Act 1981
Copyright Act 1956
Coroners Act 1887
County Courts Act 1959
Criminal Damage Act 1971
Criminal Justice Act 1967
Criminal Justice Act 1982
Deer Act 1963
Deer Act 1980
Exchange Control Act 1947
Forgery Act 1913
Highways Act 1959
Hijacking Act 1971
Immigration Act 1971
Indecent Advertisements Act 1889
Insolvency Act 1976
Legal Aid Act 1974
Ministry of Social Security Act 1966
Motor Vehicles (Wearing of Seat Belts in Rear Seats by Adults) Regulations 1991
Motor Vehicles (Wearing of Seat Belts) Regulations 1982
Post Office (Protection) Act 1884
Post Office Act 1969
Powers of Criminal Courts Act 1973
Prevention of Fraud (Investments) Act 1958
Prevention of Terrorism (Temporary Provisions) Act 1976
Prevention of Terrorism (Temporary Provisions) Act 1984
Protection of Aircraft Act 1973
Representation of the People Act 1949
Road Traffic Act 1972
Social Security Act 1986
Wireless Telegraphy Act 1949

Produced by the Research Development and Statistics Directorate, Home Office

This document is available only in Adobe Portable Document Format (**PDF**) through the RDS website

Home Office
Research, Development and Statistics Directorate
Communication Development Unit
Room 275
50 Queen Anne's Gate
London SW1H 9AT

Tel: 020 7273 2084 (answerphone outside of office hours)
Fax: 020 7222 0211
Email: publications.rds@homeoffice.gsi.gov.uk

ISBN 1 84082 856 0