Evaluation of a Database for Tracking Cases of Child Sexual Abuse

Cate Bailey , Martine Powell and Gennady N. Baksheev

Deakin University, Geelong, Australia, Centre for Investigative Interviewing, School of Psychology

Administrative databases are used by criminal justice professionals to guide specialist responses to crimes of child sexual abuse. Assumptions might be made that the database will be accurate, contemporaneous, complete, and meaningful; however, this may not be the case. The main aim of the current study was to critically evaluate a database used by practitioners for tracking cases of child sexual abuse, in order to identify evidence that may justify investment in improved data gathering and centralised information management systems. Three data quality dimensions were examined: (1) completeness, measured as data that were not missing and were of adequate breadth and depth, (2) accuracy, namely that the data are correct, and (3) believability, where the data may be regarded as credible or plausible. Results indicated that data quality was of concern for all three dimensions, with missing and inaccurate data found across a range of variables, and issues with believability found on two variables. The implications of these results for development of new data documentation methods are discussed.

Key words: administrative databases; case tracking; child sexual abuse; data quality; evaluation; information management; police.

Case attrition is significantly higher for reported instances of child sexual abuse than for other indictable offences (Victorian Law Reform Commission, 2004). Only around 8% of reported child sexual abuse cases actually result in a conviction (Fitzgerald, 2006). Attrition has been defined as the premature exit of a case from the system, and is a common, challenging, and persistent issue (Christensen, Sharman, & Powell, 2015). Because of this high attrition rate, gaining a better understanding of where cases leave the criminal justice system has been acknowledged as an urgent priority by government agencies (Community Development and Justice Standing Committee, 2008; New South Wales Ombudsman, 2012). Increased knowledge of where cases leave the system could

help policy-makers improve outcomes for children who have been sexually abused.

One way to investigate issues of attrition is to use administrative databases to track cases. Administrative databases are also important tools for case management and for managing legal requirements, and many critical decisions may be based on such data. It is essential that these decisions are based on the most accurate data available. Assumptions might be made that an administrative database will be accurate, contemporaneous, complete, and meaningful; this may, however, not be the case (Leach, Baksheev, & Powell, 2015).

Despite the importance of these databases, little is known about their possible limitations, and high-quality evaluations have

been scarce (Byrne, Regan, & Howard, 2005). Some of the data quality concerns raised have related to consistency in data recording, the identification of relevant cases, and missing data (Leach et al., 2015). Furthermore, when data have been collected for agency purposes, recording practices may have changed over time, and definitions of variables may be inconsistent between practitioners, units, or departments.

Improving data-recording practices has been the focus of increased attention across health, justice, and other government agencies (New South Wales Ombudsman, 2012; New Zealand Government, 2012; Victorian Law Reform Commission, 2004). These reports consistently highlight the need for rigorous and reliable information; and they have recommended the establishment of centralised information management systems. Centralised information management systems can facilitate improved data collection by housing the database on a centralised system, with data entry and access being available from different sites, as required. A system of this kind would have a graphical user interface that would restrict data entry to desired information types, through the use of drop-down menus and clearly defined variable entry points. This would ensure that there were fewer errors and inconsistencies, and better data security and control.

The main aim of the current study was to critically evaluate a database used by practitioners for tracking child sexual abuse, in order to identify possible evidence that may justify investment in improved data-gathering and centralised information management systems. The database used in this current study was ideal for the purposes of this evaluation, as it was a large database containing cases of child sexual abuse that had been collated by different personnel. Three aspects of data quality were analysed: (1) completeness, measured as data that were not missing and were of adequate breadth and depth, (2) accuracy, namely that the data

are correct, and (3) believability, where the data may be regarded as credible or plausible (Pipino, Lee, & Wang, 2002).

Method

A de-identified spreadsheet was sourced from a joint police and child protection service in Australia (Target database). Victims of child abuse and neglect were referred to this agency for investigation of the alleged offence; the agency was staffed by members of the police force and child protection systems. This specialist team assessed all new referrals of child sexual abuse and serious physical abuse, as well as commencing child protection and investigative actions to ensure the safety of the child and the preservation of evidence. As such, the data utilised in this study were collected in the course of normal working practices, and in line with policies and procedures of the agency. Data were kept on approximately ten separate Excel spreadsheets in different areas of the agency and were entered without restrictions by practitioners as a combination of text and numerical entries. The collated database contained information on victims referred to the agency from 2009 to 2011, comprising approximately 3000 cases per year. The database contained: (1) general information about the alleged offence (e.g., abuse type and suburb of the alleged offence), (2) information about the victim (e.g., age at the time of the incident, gender, ethnicity), (3) information about the person of interest/alleged offender (e.g., age at time of incident, gender, relationship to child), and miscellaneous information. This database was provided to the research team in de-identified format, thereby maintaining confidentiality and anonymity of the victims and alleged offenders. Data in the Target database were selected for this study from sexual abuse cases in the year 2011, so as to be comparable to the Verification database. A total of 2267 cases were retained, of which 76.9% were female victims, and 22.4% were male victims.

Verification Method

An appropriate verification method was an important consideration in this study. The comparison database used in the current study was a purpose-built database manually extracted from a police information management system. These data were considered to be more accurate than the Target database, as there were many opportunities to correct the data as new evidence became available in each case. In order to construct this Verification database, searches were conducted in the police information management system by setting parameters for date and type of offence, and collating the relevant case identification numbers. Final case documentation for each case was read in its entirety, and the relevant data were manually extracted for child, offender, and offence variables. (Unfortunately, case location data were not collected during this process.) All cases were de-identified. The database contained a total of 853 cases, of which 78.5% were female victims, and 21.5% were male victims.

The Target database was constructed in the course of normal business. It was designed to be used by many different

professionals, it had free flow access for inputting data, and data quality relied on the integrity and effort of the professionals who entered the data. These professionals were often juggling many tasks in an environment of competing priorities, of which data collection was only one. The Verification database accessed data that were not practical to obtain on a day-to-day basis. It was manually extracted from a large administration system by a team of research assistants for research purposes.

Data were considered to be missing if a variable was left blank or was coded as 'unknown' or 'unspecified' (Sanders et al., 2012). Data were considered incorrect if there was a spelling mistake, incorrect unit of measurement, inconsistent capitalisation, more than one piece of information in the cell, or inconsistent format of data. Believability was determined by comparing variables in the Target database with a database established as an acceptable standard, in this case the Verification database. The study was approved by the University's Human Research Ethics Committee and the ethics committee attached to the agency.

Table 1. Numbers and percentages of missing data in the Target and Verification databases.

Variable	Target (<i>N</i> = 2267)		Verification (<i>N</i> = 853)	
	#	%	#	%
Victim's gender	14	0.6	0	0.0
Victim's Indigenous status	1459	64.4	260	30.5
Victim's age	23	1.0	2	0.2
Delayed reporting ^a	1	0.0	34	4.0
Offender victim relationship	662	29.2	46	5.4
Offender gender ^b	0	0.0	0	0.0
Offender Indigenous status ^b	863	53.2	242	31.6
Offender age ^b	246	15.2	35	4.6
District office known	18	0.8	n/a	n/a
Suburb of offence: victim	26	1.1	n/a	n/a
Victim has disability	26	1.1	n/a	n/a
Child in foster care (ward of state)	26	1.1	n/a	n/a

^aReporting delayed for over 12 months.

^bWhere an offender was found.

Data Analysis

IBM Statistical Package for the Social Sciences (SPSS) version 23 was utilised to analyse data. Frequency and descriptive analyses were conducted to examine the level of missing and incorrect data for child maltreatment cases. Percentages were calculated for both missing and incorrect data, and to assess believability.

Results

Completeness

Analyses were conducted to determine levels of missing data in the Target database, as presented in Table 1. Variability in rates of missing data was high across the variables. Victims' age and gender had low rates of missing data in both databases, but rates of missing data were high for victims' Indigenous status in both databases, and particularly so in the Target database. Almost two thirds of the data points for victims' Indigenous status were missing in the Target database, and one third in the Verification database. Rates

of missing data were similar for offender Indigenous status, with just over half missing in the Target database and almost a third missing in the Verification database.

Missing data were higher in the Target database for offender–victim relationship, at almost 30%, compared to 5% in the Verification database. Missing data were also high for offender age in the Target database at 15%, compared to 5% in the Verification database. Only one variable had more missing data in the Verification database than in the Target database: delayed cases (where the incident was reported more than 12 months after the offence) were missing in 34 (4%) cases in the Verification database, but only one case in total was missing for this variable in the Target database.

Accuracy

Data quality was investigated for variables, as shown in Table 2. The Target database had compromised data on three variables. Data for the victims' age in the Target

Table 2. Numbers and percentages of inaccurate data on variables in the Target and Verification databases.

Variable	Target (<i>N</i> = 2267)		Verification (<i>N</i> = 853)	
	#	%	#	%
Victim's gender	3	0.1	0	0
Victim's Indigenous status	0	0	0	0
Victim's age	49	2.2	0	0
Delayed reporting ^a	0	0	0	0
Offender victim relationship	0	0	0	0
Offender gender ^b	21 ^b	0.9	0	0
Offender Indigenous status ^b	0	0	0	0
Offender age ^b	64 ^b	2.5	1	0
District office known	12	0.5	–	–
Suburb of offence: victim	49 ^c	2.2	–	–
Victim has disability	1	0	–	–
Child in foster care (ward of state)	0	0	–	–

^aReporting delayed for over 12 months.

^bMultiple offenders entered.

^cMostly lack of capitalisation.

database contained just over 2% of data that was inaccurate. Victims' age data should have been entered as number of years, entries such as '10 months' were interspersed with numbers that ranged up to age 81. There were ten entries with the victims' age as over 20 years. It is not possible to know whether some of these data points were entered erroneously, or whether they were entered in months rather than in years. Offenders' age contained 3.9% of cases where multiple entries had been made into the same cell for multiple offenders. There were three minor errors in child gender and errors in 1.3% of cases for offender gender, where multiple entries had been made in the data cell for multiple offenders. Data quality issues were apparent for the variable of the suburb where the offence had taken place. These were mainly spelling or typing errors. Similar errors were found in the district variable.

Believability

To assess believability, means or percentages of the comparable variables in the two databases were calculated, as shown in Table 3. Means or percentages for each variable were compared across the two datasets to determine whether there may be concerns about whether the data could be regarded as true

and credible (Pipino et al., 2002). Age and gender were similar between the two databases for both children and offenders. The Target database recorded very few cases where reporting was delayed for over 12 months, compared to 16% in the Verification database, indicating that this variable was most probably not recorded accurately in the Target database. The percentage of children identifying as Indigenous, however, was 12% higher in the Target compared to Verification databases for children and over 5% higher for offenders.

Discussion

This study has revealed some of the limitations of an administrative database used for recording child sexual abuse data, and the importance of evaluating the accuracy of this type of data. While the Verification database was relatively accurate, this type of database is expensive to produce, as the data were collated manually from police records, and would not be available for use in a timely manner. Key findings from this analysis suggest that the call for centralised information management systems, with clearly defined variables restricted by drop-down menus, would be beneficial for improving data accuracy in a database such as the Target database in this study.

Table 3. Comparative demographics for the two databases, where data were known.

	Target			Verification		
	<i>M</i>	<i>SD</i>	%	<i>M</i>	<i>SD</i>	%
Age	10.77	4.97		10.86	4.12	
Female			77.6			78.5
Indigenous child			34.5			22.1
Delayed reporting ¹			0.7			16.0
Familial			44.9			38.9
Offender age	29.88	16.02		31.61	16.58	
Offender male			94.4			96.3
Offender indigenous			24.9			19.5

¹Reporting delayed for over 12 months.

Missing data were found in the Target database in the following variables: relationship between offender and child, child and offender Indigenous status, and offender age. The high level of missing data in both databases for Indigenous status for both child and offender was of particular concern, with 34% and 22%, respectively, more missing data in the Target database than in the Verification database. In a centralised information management system, data entry could be directed such that each variable must be entered before the data entry process could proceed, thereby reducing missing data. Furthermore, when data were unavailable, they could be coded as such, so that the reasons for the data being missing would be clear.

With regard to data accuracy, issues were mainly focused on the variables for children's and offenders' ages. These issues could be resolved by the use of drop-down menus contained in a centralised information management system. The use of drop-down menus would restrict data options to reduce data inaccuracy.

Believability was found to be a concern in two of the variables in the Target database. Firstly, although the variable for delayed cases in the Target database had low rates of missing data and low rates of inaccurately entered data, a comparison between the results for this variable in the two databases was concerning. Very few data points were identified as delayed in the Target database, whereas 16% of the data in the Verification database was identified as such. This finding is a reminder that all data variables need to be inspected closely for believability (defined as the degree to which data are accepted as true and credible) either from the individual's assessment or via a comparison with a commonly accepted standard, in this case the Verification database.

A further concern regarding believability was revealed when comparing results for child and offender Indigenous status between the two samples. For both children and offenders, the Target database had a

significantly higher percentage of cases identified as Indigenous than the Verification database. This suggested that these data may not be missing randomly, and thus it would be ill-advised to estimate numbers of Indigenous cases from the available data. This problem was restricted to the Indigenous status variables, as results for other variables were comparable.

In summary, a centralised information management system would be more accurate and reliable than the methods used in the Target database, where data were entered unrestricted onto separate spreadsheets. In a centralised information management system, drop-down menus rather than writing free-flow would ensure that data would be accurate, spelled correctly, and meaningful. Requiring database users to complete every variable as they enter data, as well as marking those variables that were not relevant with 'not applicable', would assist in decreasing the extent of missing information. Clear descriptions of the data variable would also avoid ambiguities as to the meaning of particular variables, thereby improving believability. Further, data entered using a centralised information management system would be up-to-date and available in a timely manner for investigative and evaluation purposes. This system could potentially be compatible with administrative systems across agencies, such as court systems, further improving access to important outcome data.

This study has analysed a number of different dimensions of data quality in an administrative database designed to record cases of child sexual abuse. The findings have shown that there were missing data on a range of variables and indicated some further concerns with data quality. There was a demonstrated need for accurate and timely information on cases of child sexual abuse, in order to improve outcomes for these children. Better systems for collecting data, such as centralised information management systems, are strongly recommended.

Acknowledgements

The authors would like to thank the organisations that provided the database, and in particular the data analyst who compiled the main database. Thanks also to Mairi Benson and Chelsea Leach for compiling the verification database, and Bronwen Manger for editorial assistance.

Disclosure Statement

No potential conflict of interest was reported by the authors.

ORCID

Cate Bailey <http://orcid.org/0000-0001-5030-430X>

Martine Powell <http://orcid.org/0000-0001-5092-1308>

References

- Byrne, N., Regan, C., & Howard, L. (2005). Administrative registers in psychiatric research: A systematic review of validity studies. *Acta Psychiatrica Scandinavica*, 112(6), 409–414. Retrieved from <http://doi.org/10.1111/j.1600-0447.2005.00663.x>
- Christensen, L., Sharman, S., & Powell, M. (2015). Professionals' views on child sexual abuse attrition rates. *Psychiatry, Psychology and Law*, 22(4), 542–558. Retrieved from <http://doi.org/10.1080/13218719.2014.960036>
- Community Development and Justice Standing Committee. (2008). *Inquiry into the prosecution of assaults and sexual offences*. Perth: Legislative Assembly, Parliament of Western Australia.
- Fitzgerald, J. (2006). The attrition of sexual offences from the New South Wales criminal justice system. *BOCSAR NSW Crime and Justice Bulletins*, (92), 12.
- Leach, C., Baksheev, G. N., & Powell, M. (2015). Child sexual abuse research: Challenges of case tracking through administrative databases. *Psychiatry, Psychology and Law*, 22(6), 912–919. Retrieved from <http://doi.org/10.1080/13218719.2015.1019333>
- New South Wales Ombudsman. (2012). *Responding to child sexual assault in Aboriginal communities*. Sydney. Retrieved from http://www.ombo.nsw.gov.au/__data/assets/pdf_file/0005/7961/ACSA-report-web1.pdf
- New Zealand Government. (2012). *The white paper for vulnerable children* (Vol. I). Retrieved from <http://www.beehive.govt.nz/feature/white-paper-vulnerable-children>
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211–218. Retrieved from <http://doi.org/10.1145/505999.506010>
- Sanders, C. M., Saltzstein, S. L., Schultzel, M. M., Nguyen, D. H., Stafford, H. S., & Sadler, G. R. (2012). Understanding the limits of large datasets. *Journal of Cancer Education*, 27, 664–669. Retrieved from <http://link.springer.com/article/10.1007/s13187-012-0383-7/fulltext.html>
- Victorian Law Reform Commission. (2004). *Sexual offences: Final report*. Melbourne. Retrieved from <http://www.lawreform.vic.gov.au/projects/sexual-offences/sexual-offences-final-report>