Annex on data and statistical methods

1. Introduction

1. Human rights incidents are complex. An eyewitness or victim may report one or many victims, who may each have suffered one or many violations. Each violation may also involve one or many perpetrators. Hence, the interactions between different persons in thousands of these types of incidents require careful empirical methods of identification and aggregation in order to facilitate valid and reliable quantitative analysis.

2. To assure the quality of its data, the Commission instituted several processes. This methodological appendix presents the data and methods from which the Commission's statistical results are derived.

3. The Appendix is divided into six main sections. Section 1 provides an outline of the relevance of empirical data analysis to the Commission's mandate. Section 2 provides detailed descriptions of the different datasets which were used in the Commission's statistical analysis. Section 3 describes the data editing, cleaning and name normalisation techniques which were applied to the data. Section 4 presents the recording accounting tabulations at different stages of the data conversion process. Section 5 presents the various de-duplication and record linkage techniques which were used to match multiple reports of the same individual victim. Section 6 documents the data processing which was used to account for multiple reports of groups of anonymous victims. Finally, section 7 presents the statistical estimation techniques which were used to derive total estimates of the magnitude and pattern of fatal violations and displacements during the Commission's reference period.

Relevance of empirical data analysis to the Commission's mandate

4. The Human Rights Data Analysis Group (HRDAG) helped the Commission to collect and analyse human rights violation data relevant to the mandate period of the Commission, 1974-1999.[1] This Appendix explains how the data were organised and processed.

5. The Commission required an information management system to manage and structure the data needed to answer the issues outlined in its mandate. Specifically, the Commission's information management system had to supply information about past human rights violations which would subsequently provide:

1. Descriptive statistical analyses of general patterns and trends of violations in order to describe the "nature" of human rights violations (the types of violations which were committed).[I]
2. Statistical projections of total violations to establish the "extent" of human rights violations (the total number of violations which were committed).[II]
3. Statistical hypothesis testing of the regularity of certain violations in order to investigate whether certain patterns of violation constituted "a systematic pattern of abuse".[III]
4. Case-level analysis by basic filing and searching of the database in order to describe the "antecedents, circumstances, factors, context, motives and perspectives" that led to large-scale violations.[IV]

[1] HRDAG is a division of Benetech Inc in Palo Alto, California, USA. HRDAG staff include statisticians, computer programmers, and record linkage experts. HRDAG team members have worked in large-scale human rights documentation and analysis projects on five continents, in more than a dozen countries over the past 20 years. HRDAG has worked with official truth commissions in Haiti, South Africa, Guatemala, Peru, Ghana and Sierra Leone; with the International Criminal Tribunal for the Former Yugoslavia; and with non-governmental human rights groups in El Salvador, Cambodia, Guatemala, Colombia, Afghanistan, Sri Lanka and Iran. For more information see http://www.hrdag.org.

5. Structured quantitative analysis and hypothesis tests in order to investigate whether "human rights violations were the result of deliberate planning, policy or authorisation" on the part of specific parties to the conflict.[V]
6. Formal explanations of the scientific and statistical methodologies employed in order to demonstrate that the Commission's findings are based on "factual and objective information and evidence collected or received by it or placed at its disposal".[VI]

6. The Commission was aware that after suffering human rights violations a large proportion of victims and their families had lived in silence, fear and isolation, sometimes for more than 25 years. Therefore the Commission had to devise data collection and information management systems that would both produce reliable historical data and promote public participation in the truth-seeking process.

2. Data sources

7. This section sets out the characteristics of the three primary statistical databases the Commission established to undertake quantitative analysis of past human rights violations and promote reconciliation in Timor-Leste. The Human Rights Violations Database (HRVD) was a collection of narrative statements from victims, qualitative reports from Amnesty International (AI) and data collected by Fokupers, a local East Timorese NGO. The Retrospective Mortality Survey (RMS) was a random-sample household survey used to measure displacement and mortality during the Commission's mandate period. The Graveyard Census Database (GCD) was a comprehensive census of public graveyards in each of the 13 districts of Timor-Leste.

8. The combined data from all three of the Commission's data streams—the HRVD, the RMS and the GCD—were used to make independent demographic estimates of the total extent, pattern and trends of and levels of responsibility for past fatal violations in Timor-Leste.

The Human Rights Violations Database (HRVD)

9. The following sections describe the three documentation projects which were conducted to form the Commission's Human Rights Violations Database. The process of transforming qualitative information from these documentation projects into statistical data is also presented. Finally, the recording accounting from the three documentation projects is presented.

The statement-taking process of the Commission

10. In February 2003 the Commission began collecting narrative statements from individuals in all 13 districts of Timor-Leste and from East Timorese people then living in West Timor. These statements were the basis of the HRVD. The Commission established offices in each of the 13 districts to implement its mandate. A total of 7,669 relevant narrative statements were collected documenting reported human rights violations. These narratives provided extensive information on both fatal and non-fatal violations during the reference period.[2] The statement-taking process covered all 65 sub-districts in each of the 13 districts of Timor-Leste.[3] In addition to the district-level statement collection, the Commission also

[2] Commission teams collected a total of 7,824 statements. Some of these (155 statements) were not entered into the HRVD because they either did not mention violations connected to the Commission's mandate, or the violations which they mentioned were not within the Commission's reference period

[3] The Commission's district teams generally worked with communities according to local identification with sub-districts and villages and aldeias. As the Commission commenced work in early 2002, the common figure of sub-districts in Timor-Leste was 65; however, the National Statistics Office and the 2001 Timor-Leste Suco Survey report 64 sub-districts.

collected 86 <s00120> statements from East Timorese refugees and others living in West Timor, through the Commission's partnership with a coalition of West Timor-based NGOs.[4]

11. Given that the statement-giving was entirely voluntary on the part of the deponent, and based on a convenience sample, the distribution of statements across geographic locations was not uniform. As graph <g5000001> indicates, the Commission collected substantially more statements from deponents in Bobonaro and Ermera than from deponents in other districts (See section below for a detailed description of the possible factors which influenced the sampling process during the Commission's statement-taking process).

[INSERT <g5000001> about here]

12. In order to analyse this qualitative information statistically, it was coded into a FoxPro database using the design standards of the "Who Did What To Whom" data model.[VII] Although these data provide many useful insights, the Commission statement-taking process that generated them did not employ a probability-based random sample. Rather the Commission accepted statements from those willing to volunteer the information they could recall. As a result the narrative data, in isolation, cannot be assumed to be statistically representative of the overall extent and pattern of violations in Timor-Leste.

Demographic characteristics of deponents

13. Approximately 21.4% (1,642/7,669) <s00104> of all deponents in the Commission statement-taking process were women. In some communities, women did not participate in the Commission's socialisation activities as they were expected to stay at home. In addition fewer women were members of formal organisations with access to information regarding the Commission's work, and some women were uncertain or shy about coming forward to give testimony.[5]

14. The Commission received statements from adults of all ages. For both males and females, the highest number of deponents were in the 40-44 age group, as indicated in Figure <g500002>.

[Insert Figure <g500002.pdf> about here]

15. Despite the substantial difference in male/female participation rates in the Commission's statement-taking process, female deponents tended to talk about violations against themselves (relative to violations against others) in roughly the same proportion as male deponents. As Figure <tDepSexVictSexM> shows, of all the violations reported by females, 30.6% (2,939/9,605) were violations against themselves, whereas for male deponents, 35.3% (17,438/49,382) of reported violations were against themselves.

[Insert Figure <tDepSexVictSexM> about here]

16. The social, cultural and economic challenges faced by women may have limited their participation in the Commission's socialisation and statement-taking processes. However, the Commission's statistical findings are consistent with the claim that most of the victims of killings, disappearances, torture and ill-treatment were young males. By contrast, the overwhelming majority of sexual violations documented by the Commission were suffered by female victims (see Part 6: Profile of Human Rights Violations).

17. Statement-takers interviewed deponents in Tetum, Indonesian or other East Timorese languages or dialects (which are oral though not commonly written languages) and then wrote the text of the interview in Tetum or Indonesian. Statement-taking forms were

[4] The Coalition of NGOs comprised CIS (Center for Internally Displaced Persons Service), Truk-F, Lakmas (Lembaga Advokasi Kekerasan Masyarakat Sipil), Yabiku and Yayasan Peduli |ndonesia (YPI). Staff from these NGOs collected statements from East Timorese living in Belu, Kefamenanu, Soe and Kupang in West Timor between February and August 2003.

[5] CAVR, internal document: Evaluation Report of CAVR Statement Taking Process. CAVR Archive..

available in Tetum and Indonesian. Of the 7,669 <s00101> statements received by the Commission and found to be within the Commission mandate, 81.7% were in Tetum, 17.0% in Indonesian, 1.2% in other East Timorese languages, and 0.1% in a language that was not specified <s00002>. As the Commission's statement-taking forms were in Tetum and Indonesian, statements given in other languages were written by the statement-takers onto the official form in either Indonesian or Tetum before coding, data-entry and analysis of the narrative statements.

Potential sampling biases in the statement-taking process

18. As discussed below, the voluntary nature of the Commission's statement-taking process resulted in a degree of "self-selection". This "self-selection", in turn, introduced a number of factors which affected who was able to give a statement, such as:

- people who lived in remote and mountainous areas very far from where the data were being collected (such as district towns) had less chance of being in the sample than those closer to regional towns and district capitals
- people who were socially active and/or physically agile were more likely to give statements than those who were sick, elderly, disabled or traumatised
- people who were active in the local community or closely affiliated with local village, sub-district and district officials and elders were more likely to participate in the socialisation process and statement taking because these local statement-collection efforts were often organised through local village structures and officials
- people who died before the Commission was formed did not have an opportunity to tell their stories to the Commission; therefore, events that took place in the past tended to be less frequently reported than more recent events
- people with little or no access to the media and mass communications were less likely to approach the Commission, and
- people from constituencies that were hostile to the Commission were less likely to make statements.

19. In order to address sampling biases, the Commission supplemented the statement-taking process by the collection of narrative statements from Fokupers and secondary source information from Amnesty International. Furthermore, to account for biases in measurement of displacement and fatal violations, the Commission developed its Retrospective Mortality Survey which collected structured information from a random probability sample of households in Timor-Leste (See section below for a detailed presentation of the design of the sampling techniques and survey instruments which were used for the Retrospective Mortality Survey).

Amnesty International

20. Amnesty International reported on the East Timorese human rights situation during the Commission mandate period mostly by way of information gathering through underground networks in Timor-Leste and through its contacts within the East Timorese diaspora in Australia and Portugal.

21. The Commission received 322 reports and documents from Amnesty International, which were compiled between 1975 and 1999.[6]

[6] The Commission was unable to locate the following Amnesty International Reports:
ASA 21/12/83 UA 212/83 21 September
ASA 21/16/85 Disappearances
ASA 21/44/85 Unfair Trials and Possible Torture in Timor-Leste
ASA 21/22/87 Statement on ET by AI to the UN Special Committee on Decolonisation
ASA 21/23/87 ET: Releases of Political Prisoners

22. Amnesty International's qualitative reports and Urgent Actions were coded and entered into the Commission's Human Rights Violations Database using the same methods and standards as were used for the statements which were collected by the Commission. The information collected from Amnesty International describes the general human rights situation in Timor-Leste, as it was observed by the international human rights community at the time.

Fokupers

23. Forum Komunikasi Untuk Perempuan Loro Sae (Communication Forum for East Timorese Women, Fokupers) a local human rights NGO, constructed a violations database after the Popular Consultation-related violence in 1999.[7] The Fokupers database is constructed from open-ended interviews conducted by Fokupers's staff with local East Timorese women. Originally, the main purpose of the interviews was linked to the counselling work conducted by Fokupers. However, the objectives were extended to include documentation for investigation purposes by competent legal authorities, such as the UN's Serious Crimes Unit. The narrative statements were taken in the Tetum language.

24. Fokupers constructed its database to facilitate the publication of a report on violence against women. Their original database was centred on representing the biographical data of victims, the narrative events that were described, identifying the violations which occurred and perpetrators involved. In July 2004, Fokupers submitted these data to the Commission on the condition that personal identifiers of deponents, victims, or family members in the database would not be identified in the Commission Final Report. Commission staff recoded the data, based on the Commission's standardised definitions and coding scheme, so that these data could be analysed in parallel with the CAVR's Human Rights Violations Database.

Coding the qualitative sources (CAVR narrative statements, Amnesty International and Fokupers)

25. Data coding is the process of transforming unstructured narrative information on violations, victims, and perpetrators into a countable set of data elements, without discarding important information or misrepresenting the collected information.

26. In October 2003, the Commission data processing team reviewed the coding and data entry process in order to identify systematic errors and inconsistencies in the coding and data entry process. At the time, 2,473 statements had been coded and entered into the Commission's database. A random sample of 15% of statements (ie, 371 statements) in the database was taken, stratified on the district in which the statement was taken.

27. Each statement was reviewed by a coder: the coder re-coded the statement without looking at how it had been coded originally. Then the results of the two codings were compared and errors in the original coding were identified, noted and then changed. In addition, the coder would also review the database entry for this statement and identify and note any data entry errors and correct them.

28. Within the 371 reviewed statements, 416 coding errors were identified. 58% (241/416) of these errors were violation coding errors, 12% (49/416) errors associated with coding of the victim's affiliation, 10% (42/416) with the level of location specificity coded and 9% (36/416) were associated with the coding of the institutional affiliation of the perpetrator. Of the 416 coding errors identified, 70% (291/416) of these coding errors were errors of non-identification (ie, where the act was not identified as a violation or the person or location was not identified by the coder). Another 17% (71/416) of the coding errors resulted from the

---

ASA 21/14/91 AI statement to UN Special Committee on Decolonisation - Appendix I and II

ASA 21/24/91 Timor-Leste: After the massacre – Appendix 1

As a result, the Commission's statistical analysis of violations in Timor-Leste reported by Amnesty International does not include relevant acts and incidents covered in these reports.

[7] Fokupers was founded in 1997 to support victims of political violence through counselling programmes and other forms of assistance to women victims of violations, including ex-political prisoners, war widows, and wives of political prisoners. Its mandate also includes promoting women's human rights among the local population, especially East Timorese women.

coder including the act as a violation when what was described in the narrative did not met the definitions and boundary conditions of the Commission's controlled vocabulary. Finally, 13% (54/416) of the coding errors were the result of misclassification of an act into the incorrect violation category.

29. As a result of this coding review, the data processing team undertook three initiatives to minimise these errors in the future: (1) a number of revisions were made to the Commission's controlled vocabulary; (2) a training workshop in which the results of the review were presented to the coding team and extra training provided in the necessary areas; and (3) the implementation of regular group coding exercises where coders coded the same statements and reviewed the consistency of their coding decisions using both qualitative reviews and quantitative Inter-Rater Reliability (IRR) measures.[8]

30. The main types of revisions which were made to the Commission's controlled vocabulary were:

- a reduction in the number of violation categories to a more manageable list
- refinement of boundary conditions for conceptually similar violation categories (such as torture and ill-treatment)
- refocusing the controlled vocabulary to the measurement of violations only, not both the measurement of violations and the physical and psychological impact of these violations
- simplifying the definitions of violation categories and ensuring the syntax of the definition is more consistent with the specificity of information collected in the statements (for example, technical legal terms were reworded into common language or eliminated, as they did not fit the historical reality being measured)
- revision to the institutional actors list; both simplification of the list and hierarchical structuring of the institutions to reflect their structural relationships with each other.

HRVD data collection results

31. The HRVD's three combined data sources produced a database with records as shown below in Figure <HRVD_data_collection_results>. These records represented individual and groupvictims, both of which suffered fatal and non-fatal violations. Figure <HRVD_data_collection_results> shows the breakdown of the number of records collected in each database. Note that these numbers represent the data totals before cleaning where invalid and duplicate records were removed from the databases.

Table 1 - Figure <HRVD_data_collection_results>: Recording accounting matrix for the Human Rights Violations Database

	Statement Count	Individual Count	Fatal Violations	Non-Fatal Violations
CAVR statements	7779	38812	6778	31595
Amnesty International	267	547	122	631
Fokupers	423	4888	376	3983
Totals	8469	44247	7276	36209

32. Groups are records of unnamed victims that identify two or more victims. Some victims suffered multiple non-fatal violations, others suffered non-fatal violations and a fatal violation, and others suffered only a fatal violation. Consequently, violation totals do not sum to the victim count.

[8] Inter-Rater Reliability is the extent to which two or more coders agree. Inter-Rater Reliability addresses the consistency of the implementation of a coding system.

Retrospective Mortality Survey (RMS)

33. The Commission undertook a Retrospective Mortality Survey (RMS) to provide a probability-based estimate of displacement and deaths. This survey drew a stratified random sample of households, and used a structured questionnaire to collect information about deaths in the family and displacement events during the Commission's reference period. The survey enabled statistical estimates of the extent of natural mortality, famine related deaths, conflict-related deaths, and migration.

Statistical sampling used in the RMS

34. The RMS sample was based on a two-stage sample design. The first stage was a sample of all 2,336 aldeias in Timor-Leste, and the second stage was a sample of households within the selected aldeais.[9]

35. The population of households was stratified along the following variables: urban/rural, district location, and elevation.[10] Implicit stratification methods were used so that the list of aldeias was sorted by the following ranked variables: urban/city, district, and altitude, and a systematic random sample picked aldeias across each of the stratification variables.[11] A cumulative measure of size variable is created and a sampling interval is calculated as the number of clusters (144) divided by the total measure of size (180,015), which equals 1,250.1. A random number between 1 and 1,250.1 was generated (397.235) and the aldeia with a cumulative measure of size above that number was selected in the sample. 1250.1 was added repeatedly to the initial randomly generated number and aldeias were selected throughout the list in the same fashion.

36. The decision to draw a fixed number of 20 households, instead of something proportional to the size of the aldeia or some other allocation method, is primarily one of operational considerations. Selecting a fixed number of households per aldeia is one way of retaining control of the overall sample size and of having an approximately uniform distribution of workload among interviewers.

37. The Commission considered the feasibility of incorporating East Timorese respondents still displaced in West Timor into the reference population.[12] However, security, operational and data quality concerns arising from conditions in West Timor made survey implementation there difficult. Therefore, the reference population that was sampled by the Commission consisted of all households within the 13 districts of Timor-Leste.

38. It was not optimal, for both statistical and operational reasons, to allow aldeias with fewer than 20 households to be sampled. Therefore, small aldeias were combined with nearby aldeias (which were not necessarily adjacent), before sampling took place, so that the estimated number of households in a cluster (defined as an aldeia or group of aldeias) was at least 40, to reduce the chance that a sample cluster had fewer than 20 households. In practice, due to the inaccuracy of the frame, on arriving in an aldeia a field team might find that it had fewer than 20 households, either because the number of households reported in

[9] An aldeia is the smallest administrative unit in Timor-Leste. In general, an aldeia is a settlement of group of homes in a small local area. Usually, a *suco* (village) is made up of three or four aldeias, and groups of *sucos* make up a sub-district which is an administrative subset of a district. According to the 2001 Timor-Leste Suco Survey there are 13 districts, 64 sub-districts, 498 *sucos*, and 2,336 aldeias in Timor-Leste. The Commission's district teams generally worked across 65 areas considered by communities as sub-districts, as administrative boundaries took some time to be reorganised following the end of the occupation.

[10] Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling. The strata need to be mutually exclusive such that every element in the population may be assigned to only one stratum. The strata should also be collectively exhaustive, in that no population element can be excluded. Random sampling is then applied within each stratum. Stratified random sampling often improves the representativeness of the sample by reducing sampling error.

[11] The Commission used a method known as Probability Proportional to Size (in this case "size" refers to the number of households and not population, although the two are obviously correlated), a common design in surveys of this kind.

[12] Section 3.3 Regulation 2001/10 states: "The Commission may conduct all such activities that are consistent with the fulfillment of its mandate within the present Regulation."

the 1990 census was inaccurate, or because it had changed in the intervening years. For these reasons the 144 sampled aldeia clusters actually contain 165 aldeias. Operationally, this means that in these clusters, interviewers had to draw a random sample of 20 households from among the combined total number of households in the cluster.

Questionnaire design and development for the Retrospective Mortality Survey

39. The RMS questionnaire was designed to fulfill the following objectives:

- to produce estimates of total mortality in Timor-Leste between 1974 and 1999, using both survey-based estimation techniques and Multiple Systems Estimation techniques, and
- to develop survey-based analysis that estimate and describe the complicated displacement movements within Timor-Leste throughout the Commission's mandate period.

40. As a result, the questionnaire was organised into the following modules:

- a household register
- a head of household displacement register
- an adult female birth history
- an adult male/female sibling history
- an adult male/female parental history
- a general human rights violation section

41. The questionnaire[13] was reviewed by three human rights statisticians external to the Commission[14] and several subject specialists at the Commission. Through this review process, improvements were made to the layout and design of the questionnaire, and a number of terminological issues in the Indonesian and Tetum languages were identified.

42. A series of eight cognitive interviews were conducted during the questionnaire development phase. Cognitive interviewing explores the cognitive processes of the respondent. It seeks to identify difficulties and possible solutions to challenges faced by respondents in (i) comprehension of the question, (ii) retrieval from memory of relevant information, (iii) decision processes, and (iv) response processes.[15] A total of eight subjects—four in laboratory conditions and four in the field—participated in the cognitive interviewing. Significant insight was gained from the probing on respondent's date recall. In particular, cognitive processes and responses about time and date related questions indicated that often, when a respondent answered “Don't Know”, they might just not know the exact date according to the Gregorian calendar. However, their responses indicated that sometimes the timing of events were easier to recall by reference to other markers of time such as other major events, or points in the agricultural or seasonal cycle.

43. From the cognitive interviewing process, we developed structured date probes which asked the respondent to narrow event-dates into a six-month window which could be defined by major events such as holidays, or environmental/physical indicators (height of corn or other crops, rainy season or dry season). The cognitive interviewing process also indicated that temporal concepts such as “beginning”, “middle”, and “end” were not understood by all respondents, so further narrowing of the time window was not possible.

[13] See Appendix to this Annex for a copy of the survey questionnaire.
[14] Fritz Scheuren, President of the American Statistical Association, consultant to HRDAG on projects for Kosovo, Guatemala and Peru; William Seltzer, Fordham University, and Jana Asher, co-author of HRDAG reports in Kosovo, Sierra Leone and Peru.
[15] Tourangeau 1984.

44. During the cognitive and field test interviews, respondents often simply answered "Don't Know" or "into the mountains/forest" as the place to which they were displaced. As a result of the cognitive interviewing, a careful set of probes was created to elicit more detailed descriptions of the places where people were displaced.

45. After peer-review and the cognitive interviewing process, the finalised questionnaire was then translated and back-translated into both Indonesian and Tetum. The questionnaire was then field tested for 5 days in aldeias within Dili, which were not part of the sample. As a result of this field test, a few further question-sequencing, grammatical, and syntactical improvements were made.

Survey implementation and fieldwork

46. Within each sampled household, the head of household responded to both the household register (in which all residents of the household were logged) and the displacement section. An adult female was then randomly selected from the female adult population of the household to answer the adult female birth history module.

47. Before leaving each aldeia, all questionnaires were checked by field supervisors to identify and correct any mistakes or inconsistencies in the completed questionnaires. Two field coordinators accompanied the team of 22 survey enumerators into the field.

48. Twelve aldeias which had been included in the sample were not able to be visited by the enumeration team. The team was unable to conduct interviews in these 12 aldeias due to security concerns at the time. Figure <RMS_non_sampled_aldeias> lists the 12 aldeias that were not enumerated.

Figure <RMS_non_sampled_aldeias>: Aldeias in RMS sampling plan not visited by enumeration team

District	Sub-district	Suco	Aldeia
Alieu	Remexio	Liurai	Coto Mori
Baucau	Fatumaca	Samalari	Osso Luga
Baucau	Laga	Samalari	Soru Gua
Bobonaro	Atabae	Atabae	Heleso
Bobonaro	Bobonaro	Tapo	Tapo
Covalima	Fohorem	Datorua	Fatulidun
Lautém	Iliomar	Ailebere	Heitali
Lautém	Lospalos	Fuiluro	Kuluhun
Liquiça	Bazartete	Fahilebo	Fatu Neso
Oecusse	Passabe	Abani	Na Nos
Viqueque	Ossu	Uaibobo	Sogau
Viqueque	Uatu-Lari	Matahoi	Loko Loko

49. Furthermore, in some aldeias less than 10 households were enumerated resulting in some additional non-response. Overall, of the 1,440 households in the sampling frame, there was a 3.1% (44/1,440) non-response rate. Given the low non-response-rate, no explicit statistical imputation was performed to control for non-response in the survey.

Graveyard Census Database (GCD)

50. In order to develop baseline mortality data for Timor-Leste, the Commission undertook a census of public graveyards in the 13 districts of Timor-Leste. Through this process available information about names, dates of birth, dates of death, and religion was collected. Gravestones that lacked such information were also enumerated and their size was noted.[16] By collecting this information, the Commission created a *de facto* vital registration system for the East Timorese population. That is, the GCD created a baseline listing of some, perhaps even most deaths, which could be used for mortality analysis beyond this project.

[16] The size of an unmarked gravestone can be used as a proxy indicator of whether the deceased was a child or an adult.

GCD data collection

51. To facilitate the Commission's census of public graveyards in the country, a list of all known public graveyards in Timor-Leste was enumerated by CAVR field staff in consultation with village-level officials at the suco (village) level, and where possible, the aldeia level. A "public graveyard" in this study was defined as a location which is reserved exclusively for burial of deceased persons. This definition includes communal burial sites which are established on public land or land owned by a religious institution. However, it excludes family graves located on private property.

52. The GCD data was collected by two separate data collection teams. The first team collected 128,751 records from 803 cemeteries, which were entered into an series of Excel spreadsheets. The first team covered portions of all 13 districts, but only Dili was covered completely. A second team went into all districts, except Dili, to finish the graveyard survey. They collected 153,057 additional records from 1,779 cemeteries. The second team used a FoxPro database for their data entry.

53. The Commission enumeration teams documented all gravestones within public graveyards—both marked and unmarked. A marked grave was defined as having a physical structure which memorialised a person's life, with legible markings in English, Indonesian, Tetum, or Portuguese.[17] On all enumerated marked gravestones, the following information was coded if on the gravestone: full name, date of birth and date of death. Unmarked gravestones were typically small simple crosses or other burial markers, without name or date information for the deceased. Enumerators were asked to note information about the religion, type-of-material and grave size, if it was discernible from the gravestone, for both marked and unmarked gravestones.

3. Methodological description of data editing, cleaning and name normalisation techniques

54. Each of the three databases used by the Commission required data editing, cleaning, and name normalising techniques in order for the data to be compared and linked between databases. Several months were spent reviewing the data for obvious typographical or spelling errors, and a random sample review was conducted to ensure data accuracy. Technical problems occurred in converting data from one database structure to another, and these were also identified and corrected.

Database cleaning and editing

55. The data processing team carried out a complete check (and corrections where required) of all HRVD records with:

[17] Due to a lack of resources, the Commission were unable to enumerate Chinese graveyards.

- missing district/sub-district information
- implausible violation date information (eg day = 42, month =13)
- records where the violation occurred before the victim's birth date
- records where the violation occurred after the victim's death date
- statements where the deponent was coded as a victim of a fatal violation
- records where the victim's age was coded as 0 or as a negative number
- records where the victim's age was coded as greater than 75
- records where there was no violation code recorded
- records where there was no victim recorded for a coded violation
- records where there was no (individual/institutional) perpetrator assigned to a coded violation.

56. In addition to the complete quick-checks described above, the coding team also did checks of a simple random sample of records of fatal violations, detentions, torture, ill-treatment, forced recruitment, sexually-based violations and displacements. The objective of the quick checks was to identify whether there were any systematic errors in assigning affiliations of victims and institutional perpetrator responsibility. One major inconsistency was identified - namely where victim affiliation was not assigned to all victims of a violation or violations which happened in the same act or acts closely linked in time. These records were identified, and appropriate rules were applied to correctly assign victim affiliation across violations in the same act or proximate acts for the same actor.

Date editing and cleaning

57. Records that had obvious errors, such as dates of birth, violation, or death that were subsequent to the current date were examined and corrected. This was especially common in the GCD database where the grave markers were so small that full four-digit year dates could not be written out. The data entry system defaulted the two-digit year dates, which should have been in the 1900s, as the 2000s. Enumerators from different teams sometimes used different date coding standards. Some used the European standard DD-MM-YYYY, some the US standard MM-DD-YYYY", some a YYYY-MM-DD format, or variations of these using a two-digit year. Furthermore, sometimes different separators were used between years, months and days – including "/", ".", and "-". As a result, all date formats across all three datasets were mapped to the standardised format, YYYYMMDD.

58. If the DOB was after the DOD, the dates were swapped. Two types of errors which caused dates with months greater than 12 or days greater the 31 were also identified and examined. The Commission discerned that some errors were caused by variations of the spreadsheet date format settings on the data entry computers.

59. Other errors were obviously typographical. Records from the HRVD and the RMS were corrected by reviewing the original paper material and applying corrections to the database. For the GCD database there was not enough time to hand review the source, so if the error was not easily corrected, the values in that part of the date field (month or day) were left blank.

Age editing and cleaning

60. Age data were examined for possible typographical errors, for example, people over the age of 100. The sources for these records were reviewed to verify the data and corrections made as necessary. Where DOB and DOD information was known, the age was derived. The GCD age value was calculated and a new field generated to facilitate easier matching.

Violation and relationship codes editing and cleaning

61. Reviews were conducted of the violation codes and relationship codes within the HRVD and RMS identified codes that were not valid or conflicted with other data within an individual record (for example, a female being coded as a father). The paper source files for these records were reviewed and the corrections made to the database.

Geographic location code editing and cleaning

62. The geographic location data collected for the RMS and HRVD databases was coded to the East Timorese geocode standards established by the government and approved for use by the the Commission. Locations were divided into four administrative levels—District, Sub-district, Suco (Village), and Aldeia. For those locations that were outside of Timor-Leste, codes for West Timor and Java were created and when the location was not known, they were marked to a separate code for unknown. Each cemetery was given a unique code, called an "id", in order to differentiate between cemeteries in the same geographic area.

63. The GCD was not collected with the East Timorese geographic code standard, so it was translated to the standard codes.

GCD deduplication of cemeteries and graves

64. Several factors led to duplicate records of graves and graveyards in the GCD database.

- Different data collection teams inadvertently covered the same cemetery. Many cemeteries did not have posted names, making it hard to identify duplicated records strictly by cemetery name.
- The exact suco (village) and aldeia location was often hard to determine in some rural areas. Even if the cemetery had the same name, it might be coded to a different geographic location. Additionally, many cemeteries shared the same name (Santa Cruz being the most common name), which meant that cemetery name alone was not enough to determine duplicate cemeteries coded to different geographic codes.
- Many of the cemeteries in Timor-Leste were not organised linearly. This sometimes led to the team of enumerators crossing over the same gravestone, recording it more than once.
- Because of the massive amount of paper files required to gather all these data, it was possible that there were data entry duplications.

65. It was possible to find linkages between cemetery id's by examining the names of the deceased, cemetery locations, cemetery names, and complete dates of birth and dates of death after matching.[18] When rows of duplicates were found, one of the cemeteries was dropped from the dataset used for analysis. While it is common for people to have the same forename and surname, and possibly the same date of death, it is highly unlikely that they would have both the same dates of birth and death. Therefore, any records that had the same forename, surname, date of birth, and date of death were considered duplicates, and only one record was kept in the database for analysis.

66. The goal of the GCD de-duplication process was to ensure that the deceased were counted only once. It was initially thought that during the forced displacements people may have initially been buried where they died, with the body later retrieved by the family and interred at a cemetery in their home aldeia. It was also thought that if the body was not recovered, that a memorial marker in the local cemetery might be erected. While this may have occurred, careful review of the data did not reveal reburial or post-hoc marking with a memorial stone to be common practice. Furthermore, when the bodies were recovered, the

[18] A complete record is defined as having day, month and year for both DOB and DOD.

first marker would likely have been removed or relocated with the body, thus preventing over-counting. People who were never buried, or who not were buried in public cemeteries, fall outside of the GCD. In order to account for the deaths that are missing from the HRVD testimonies, the RMS interviews, and the GCD grave data, we conducted multiple-system estimation of the total deaths. This analysis is described below.

Name-cleaning processes

67. The names of persons in the Commission data needed to be addressed in two ways. First, the names needed to be parsed into three categories—first, middle/nick and last—names. Once this was complete, name canonicalisation was required to facilitate record linkage. Canonicalisation is a process of reducing each name to the simplest and most significant form possible, without loss of generality.

68. Person names contained a significant amount of variation in the spellings, in apportionment to the three name fields, and in punctuation. Name variation has many causes. In open-ended narrative statements, such as the HRVD, the deponent may be a close relative, friend, neighbour or distant acquaintance of the victim, and he or she may or may not know how to spell the names of the reported victim. Transcription by the statement-taker may involve application of additional spelling and punctuation rules and even incorporate spelling errors. Similarly spelling and punctuation transformations may take place at the data coding and data-entry stages.

Name parsing

69. To address the significant variation in how names were apportioned to the three name fields; first, last, middle/nickname, the names were parsed according to strict rules. HRDAG decided to divide the names using the "first" first name for *first*, and the "last" last name as *last*, and all other names placed into the *middle/nickname* field. Additionally the prepositions (for example, de, da, do, dos) were dropped from the name fields as their use was inconsistent in the data.

70. For example, the Portuguese name Maria Luisa da Costa da Silva may be been entered into the database as shown in Figure <name_parsing_portuguese_names>:

Figure <name_parsing_portuguese_names>: Hypothtical ways in which a given Portuguese name might be initially represented in the database

First	Middle/Nickname	Last
MARIA LUISA		DA COSTA DA SILVA
MARIA	LUISA	DA COSTA DA SILVA
MARIA LUISA	DA COSTA	DA SILVA
MARIA	LUISA DA COSTA	DA SILVA
MARIA LUISA		SILVA

71. The name parsing process would have standardised these names so that the first name was Maria while the last name would simply be Silva. All other names, less the prepositions, were moved into the middle/nick fields.

72. The indigenous East Timorese name Mau Bere may have been entered as:

Figure <name_parsing_animist_names>: Hypothtical ways in which a given Animist name might be initially represented in the database

First	Middle/Nickname	Last
MAU BERE		
MAUBERE		
MAU		BERE
		MAUBERE

73. The name parsing in this case would place Mau in the first name field and Bere in the last name field.

Name canonicalisation

74. Name canonicalisation was applied to the first and last name fields of the records after parsing to facilitate easier matching, especially the automated algorithms for record linkage. Spelling variants for names were distilled into a single representative form for each name. For example, the following spelling variations were canonicalised to AGUSTINO:

- AGUSTINUHO
- AAGUSTINO
- AGUSTIO
- AGUSTINUS
- AUGUSTINHO
- AGUSTINO
- AGUSTINU
- AGUSTONIO
- AGUSRINO
- AGUSTINHO
- AGUSTIMHO
- AGSSTINHO
- AGSTINHO
- AUGUSTINO
- AGOSTINHO
- AGUASTINHO
- ANTGOSTINHO
- AGUSTINHU
- AGOTINHO
- AGOSTINO

75. The indigenous East Timorese names were harder to canonicalise as they were generally four or five characters long and some records that appeared to be spelling variations were in fact distinctly different names. Conservative canonicalisation was applied to the indigenous East Timorese names and then tested with sample linkage of animist records looking at date, age and place information to determine additional canonicals to apply.

76. After several parses over the names to canonicalise them, a new field was generated with the name spelled out in reverse order. Then, by sorting on this new field, we were able to find additional names to be canonicalised to a single form as beginning letters could vary depending on pronounciation, but the ending syllable was likely to be the same. This process proved to be very helpful in finding additional canonicals.

77. There were Chinese, Indonesian (Muslim), and Anglo-Saxon names in the databases, as well as Portuguese names and indigenous East Timorese names. The relatively few numbers of Chinese, Indonesian and Anglo-Saxon names did not require special handling. East Timorese staff, in Timor, identified whether names were indigenous for the application of matching rules and algorithms, because indigenous East Timorese names are not always sex-specific.

78. The HRVD and RMS databases are smaller then the GCD, so we canonicalised them first. We then applied the lists of name canonicals to the GCD. The resulting names were then reviewed to identify additional canonicals.

79. During the canonicalisation process, some letters in names were found to be interchangeable with each other, most commonly with the Portuguese names. The letters S, J, G, and Z were often interchanged with each other in names. Also, the letters V, U, W, and B were often interchangeable. Less frequently, the letters H and E were interchanged, or simply dropped, for example Helder/Elder, Henrique/Enrique. An example of interchangeables would be for the name Virginia, which could be spelled with a B or V. For example, spelling variations found for the canonical VIRGINIA included BIRGINIA, BERGINA.

80. Names that began with these letters were compared to each other to assist in the canonicalisation process. Where names had more than one interchangeable or the interchangeable letter was in the middle or end of a name, it was very difficult to find potential canonicals. Therefore, a program was written that generated a list of names where combinations of interchangeable letters matched another canonical name. The record linkage expert reviewed these combinations to determine if they should be canonicalised or were distinctly unique names. Where there were additional canonicals due to interchangeables, they preferred letter for the canonical was S (for S, J, G, and Z), V (for V, U, W, B), and H (for H and E).

81. Additionally, in the canonical process, it was noted that ANJU and ANJO were often cited as the first name or the only name for a record. *Anju* is commonly used to refer to a dead infant and was found often in the GCD records when a child died before being baptised and therefore was not given a Christian name. Records with *ANJU* and a last name were used for the matching process because there was some identifying data, but records with only *ANJU* were too ambiguous to make reasonable judgments for matching.

Sex and ethnicity coding

82. During the canonicalisation process, the Portuguese first names were reviewed with the frequency of the sex codings male, female and unknown displayed.[19] Sex codings that were obviously incorrect were corrected. As with most Latin names, those that end with A generally are female and those ending with O (or U) are usually male. Where first names ended in letters other than A, O or U, the frequency between male codings and female codings were examined and when the disparity was great, indicating that a few records were miscoded during data entry, corrections were made to the database.

4. Data conversion

83. In order to expedite all the data processing steps associated with matching of duplicated records, each dataset was transferred from its original FoxPro or Excel database platform, to our Analzyer database platform.[20] The FoxPro database schema was first duplicated in PostgreSQL for importing into Analzyer. The relational database structures for the HRVD and RMS data were maintained in Analzyer.

84. Figure "<Record_Account_Pre_Post_Cleaning>" shows the total number of records from each dataset that were imported into Analzyer. Note that these totals reflect data cleaning changes which resulted in the dropping of duplicate and invalid records. [21]

Table 2 - Figure <Record_Account_Pre_Post_Cleaning>: Total record count by database Pre & Post Data Cleaning

Database	Pre-Clean	Post Clean
HRVD	41,456	37,651
RMS	4,883	4,619
GCD	195,468	149,087

[19] Frequency is a count of the instances a name or code appears in a particular data field. Values with very low frequencies can reveal potential errors or misspellings in the data.

[20] Analzyer is a free and open source application used to collect, maintain, and analyse information about largescale human rights violations. For more information about Analzyer, see HDRAG website at http://www.hrdag.org/resources/data_software.shtml.

[21] Invalid records were records for which the reported violation occurred outside of the CAVR's mandate period (i.e. not between 25 April 1974 and 25 October 1999) or records for which no date information was recorded for the associated violation.

5. Record linkage overview

85. Individuals reported in the HRVD and the RMS are sometimes reported multiple times, by different deponents and may also appear as records in the GCD. To ensure the statistical analysis controlled for duplicate reports of the same person, the data required record linkage, also known as matching. Matching was applied to two general categories of violations for study - fatal and non-fatal violations. Fatal violations included civilian killings, deaths due to deprivation, disappearances, and combatant deaths. Non-fatal violation categories included attempted civilian killing, detention, torture, rape, sexual slavery, sexual violence, ill-treatment, displacement, forced marriage, impediments to reproductive rights, unfair trial, destruction of homes, destruction of livestock, extortion, threats, forced recruitment and forced labour.

86. There were two types of matching done for the purposes of statistical estimates; intra- and inter-system matching. Intra-system matching links records that identify the same person within a single dataset, and each record can match to zero, one, or many other records within that dataset. Inter-system matching joins two or more lists of unique records from different data sources together so that a multiple systems estimation of the violations can be applied. Records matching during inter-system matching can match only to zero or one other record in each of the other datasets.

87. Due to the complexity of inter-system matching and time constraints for the work, the non-fatal data in the HRVD and RMS only had intra-system matching performed for descriptive statistics. The fatal data, which included the GCD data, was both intra- and inter-system matched as the basis for multiple systems estimate calculations. Matching was done using three methods: hand-matching, computer-generated matching, and computer-assisted matching. Each of these methods may involve more than one pass.[22]

Matching rules

88. Each individual record was compared to all other records in each dataset for possible matches and was deemed a match when a significant number of the field values match *exactly*, were in *close proximity*, or did *not conflict*. The fields used for matching were: first_name, last_name, age, sex, DOB, DOD, place_of_birth (POB), and place_of_death (POD). The middle/nickname and interview_location fields were also available for clarification purposes, but were not fields available in all three datasets, and were often sparse where they were available. While not part of the matching rules, these data were taken into consideration by the record linkage expert. However, it was not used in any computerised auto-matching.

89. The matching decisions used for the Commission data tended to over-match records.[23] Over-matching reduces the number of unique records and therefore will tend to lower the estimates. Over-matching is preferred in cases where there is uncertainty that a match is accurate, to produce conservative estimates.

Matching names

90. The first and last name fields were not always complete; some had initials or were missing either the first or last name. Attempts were made to match every record even when it was incomplete, but for fatal matching, records with neither first or last names or had initials only, were dropped from matching as there was not enough data to make reliable judgements. For non-fatal matching, attempts were made to match violations with DOB, DOD, and death location information to other records with the same values in those fields, even when there was no name or the record only had initials. Records with less complete name data relied more heavily on perfect dates and places to be matched to other records. Many

[22] A pass is a review of all the data in a dataset based on sort order or algorithm, to look for matches.
[23] Over-matching means that linkages are made between records that might not in fact be duplicates.

people could have died on the same day in the same place, and knowing which of those people to match an incomplete name to is difficult and unreliable.

Matching sex and ethnicity

91. Where the sex of the victim was known, it was only potentially matchable to records of the same sex or those with unknown sex. Records where sex was marked Unknown were matchable to records coded Male and Female, but within a matched group, the sex codes could not conflict with other records in that group.

Matching locations

92. Geographic location codes used for the CAVR data were divided into four levels: district, sub-district, suco (village), and aldeia. The GCD database was the only dataset to disaggregate location information to the aldeia level, so it was not used for matching purposes. The frequency of displacements made location information difficult for witnesses to pinpoint exactly, except in places where the violation occurred in the place where the witness currently resided or from where they originally were displaced. People may have been displaced multiple times, across multiple locations and because the conflict was spread over three decades, recall of exact locations was subject to a number of errors.

93. Additionally, the boundaries between geographic locations is affected by three factors—changes to place names and the geographic boundaries of administrative boundaries over time, the imprecision of boundaries, especially in rural areas, and potential data collection, coding, and entry errors. As a result, matches anywhere within a single district and between bordering districts was considered. Potential matches between a sub-district and suco that were closer to each other were given a higher preference as well. In studying the data closely, records that matched on a preponderance of data fields other than place provided substantiation for our judgments on location matching. Where the HRVD documented a death occurring in the same location as the interview location, it was assumed that the location information was likely to be accurate.

94. In rare cases matches were made that violated the rule for location data, but only when it was clear that the records identified the same person, and that common typographical errors accounted for the difference. When there was more than one possible match, the matching algorithm tried to match to the less-specific records in order to preserve more-specific records for later match candidates. When there was equal distribution between locations at any geographic level, the less specific location was preferred and if there one was not more less specific, than one was randomly selected to be the "rep rec".[24]

Matching dates

95. As the conflicts in Timor-Leste occurred over a long period, many respondents did not remember the exact dates and places in which events occurred. The GCD data were assumed to be more accurate for date and place information because bodies would normally be buried shortly after death, and close to the place of death. When matching on the date field, the record linkage expert would link records that were plus or minus three years from each other. The exceptions to this rule were rare, and only made when the other data fields were strong exact matches. Records with month and day data were often inaccurate in the HRVD and RMS data as memory tends to be faulty over such a long period. Therefore, more-specific dates were matched to each other where they were close, and to less-specific dates where they were not close.

[24] The “rep rec” is the record that best represents that grouping of matched records by having the most complete data. Records with the most common date or place within that group or a record with a more precise place or date are considered more complete. The more complete the data, the better each subsequent round of matching for both intra- and inter-system matching will be. Because records were being linked together and the data unique to each record preserved, as opposed to deleting duplicates, it was necessary to look at the variation within the matched records to see if the differences would significantly change the analysis.

Record level constraints

96. Matching constraints were implemented to prevent over-matching. Specifically, the following matches were not allowed:

- Records of victims from the same statement (because each statement identified unique victims who may have had the same names because of familial relationships
- Two non-fatalities could not be matched if they were reported in the same source record (because the data coding and database representation methods used prevented duplicate records from a single statement being entered into the database)
- A deponent could not match to a fatal violation
- A non-fatal record could not match to a fatal record if any dates associated with the non-fatal violations were before the fatal records DOB
- A non-fatal record could not match to a fatal record if any dates associated with the non-fatal violations were after the fatal records DOD.

Intra-system matching

97. Within a dataset a person may be identified by multiple witnesses. Intra-system matching links records that identify the same person to generate a list of unique named persons to prevent over-counting, and thus, over-estimations. Intra-system matching is very complex and difficult to perform in a database as a person can match to *n* other records in the dataset. Therefore, the data are manipulated in a spreadsheet which makes it easier to order and reorder the data in multiple ways to locate linkages that need to be made.

98. Intra-system matching a dataset before merging its records with other datasets can reveal patterns inherent in that data collection project. Some of these patterns may be systematic errors in data collection, coding or data entry, or may be the result of the structure of the data collection. The observation of patterns within each dataset allows for the investigation, and if necessary, the correction of the underlying errors.

99. The three datasets of the Commission would have been too large to do high quality data matching if combined because some of the patterns would have not have been noticeable to the human eye. That is, if all three datasets were combined into a single list, the resulting list would include more than 160,000 records. Finding matching records in a list this long would have been very difficult for a human reader.

HRVD intra-system fatal matching

100. First, intra-system matching on fatal data in the HRVD was performed to link records that described the same victim. The records were imported into a spreadsheet and sorted on first name, last name, POD, and DOD, to find records that matched.

101. As records were linked, a “rep rec” was chosen. After each sort, a matching pass was performed and the linked records within a match group hidden (but not dropped) from the outputted data file, leaving just its “rep rec”. This reduced the *noise* within the data. Noise can be defined as the non- “rep rec” records in a match group that distract the matcher from the potential relationships of the “rep rec” to other candidate matches. The smaller the list of unique records , the easier it is to see potential matches and other patterns within the data. Each subsequent pass identifies additional matches, and finally, a list of unique records is distilled from the entire dataset. A minimum of five passes are done on each dataset.

102. The 15,043 fatal records of the HRVD dataset were reduced to a list of 11,145 unique victims. All the records are then imported back into the Analzyer data matching system. The matched records were linked back to the “rep rec” for analysis when all matching was completed .

RMS intra-system fatal matching

103. The RMS intra-system fatal matching was performed in a spreadsheet after the HRVD intra-matching was completed. The RMS intra-system matching used the same fields as the HRVD intra-system matching and also looked at the source of the record. Records of fatalities collected from the same household were not allowed to match to each other as they identified unique individuals, even if they shared the same name and DOD.

104. The 4,883 fatal records of the RMS dataset were reduced to a list of 4,619 unique victims.

105. The resulting linkages of both the HRVD and RMS datasets were imported back into the Analzyer data model for use in computer-assisted and computer-generated matching, and to generate data for analysis. Information and patterns documented by the record linkage expert in the hand-matching phase was then used to generate matching rules and algorithms for the computer-assisted and computer-generated matching processes.

HRVD intra-system non-fatal matching

106. Computer algorithms were devised to clean and match non-fatal violations in the HRVD. This step is referred to as *auto-matching.* Automated matching algorithms for the non-fatal violations in HRVD were developed as time and resource limitations did not permit the use of a human record linkage expert. There were three times as many non-fatal victims as fatal victims reported in the HRVD.

107. The HRVD contained 41,546 records. The intra-system auto-matching yielded a list of 37,651 unique victims of fatal and non-fatal violations.

Auto-canonicalisation of non-fatal name values and matching

108. The first step in the auto-canonicalisation process was to build a table with the different cleaned versions of all (fatal and non-fatal) original names in the database. For the first name, the versions were *normalised*, *normalised-terse*, *first word of normalised* (called first-namefirst), and *first word of normalised-terse* (called first-namefirst-terse). The same method was applied to the last name, except the last word was used instead of the first word. Then, for each victim name of a non-fatal violation, an attempt was made to match the following combinations of the normalised non-fatal full names to all of the normalised hand-canonicalised full fatal names:

- namefirst + namelast
- namefirst-terse + namelast-terse
- first-namefirst + last-namelast
- first-namefirst-terse + last-namelast-terse

109. The matching program matched on a full set of information before trying to match on less information. This matching of non-fatal to fatal-names was only done for normalised fatal names that mapped to a unique canonical name; as the information became more terse, there were fewer and fewer "allowable" normalised names to match on (which was offset by the fact that it was easier to make the match, because the less-terse information was more resistant to coding variability and data entry errors).

110. For those full names that could not be canonicalised, the first names and last names were canonicalised independently. The order of matching first names was as follows:

- namefirst
- namefirst-terse
- first-namefirst
- first-namefirst-terse

111. A subsequent matching process was developed to follow the preliminary matching round based on the auto-cleaning and auto-matching processes. This process targeted potential matches with the non-normalised names and identified the information-density per data-field of each name record. The percentage of records that contained non-blank values for the respective data fields was as follows:

- 9% had date_birth (all of these have birth_geo1)
- 44% had birth_suco_location
- 50% had birth_subdistrict_location
- 53% had birth_district_location
- 70% had Firstname
- 94% had Sex
- 100% had Lastname (since it's a mandatory field required for matching)

112. Since the last name field was the only non-blank field for all records, it was the only field that could be used in the index blocking. Blocking looks at records where the field(s) being blocked share the same value. The blocking for the last name field was done on the first four letters of each name. The match algorithm had to be carefully calibrated: if there were many blank fields, then a closer match on the non-blank fields was required (also, matches on very common last names were given less weight).

113. There were three different kinds of “closeness” that were varied:

7. The number of letters in the name that matched (4, 8, or all)
8. The number of levels in the birth location that matched (from 1 through 3), and
9. The required-closeness of the dates (from 1/3 year to 3 years).

114. With two-thirds of the victim names auto-canonicalised, and a well-defined set of rules for required-closeness-of-match for different numbers of non-blank fields, the resulting match rate was approximately 15% (compared to about 25% for the human matched fatal-violations data).

115. A match rate of 15% for non-fatal violations seems plausible as:

- Only two thirds of the name records could be canonicalised, and
- It is usual to expect higher reporting density for fatal violations as they are more easily identifiable and easier to recall by a larger number of people in the victim's social network.

116. The automated inter-system matching on the non-fatals reduced the dataset from 44,203 records to a list of 31,568 unique victim records.

Data linkage expert review of HRVD non-fatal intra-system matches

117. The record linkage expert studied a sample of the auto-matched results to make sure there were no obvious mis-matches (ie. over-matching). No systematic pattern of over-matching was found in the review of a random sample if 10% of the matched group records. The largest group of records which were matched to each other was 20 records. A review was done of the largest groups to ensure that their match size was plausible.

118. Intra-system matching on fatal data generates a combined list of unique individuals who are all dead, even though the cause of death can vary. When intra-matching is done on non-fatal violations, a victim can suffer one or more violations, on one or more days, in one or more places. The non-fatal matching reveals the human rights violations suffered by individual victims, where a victim may have suffered other violations that may or may not have resulted in a fatality.

Inter-system matching

119. Inter-system matching links lists of unique individuals from multiple datasets and is done cumulatively in pairs or datasets. Inter-system matching is applied only to fatal data. First, inter-system matching is applied using the 11,126 intra-system matched records from HRVD to the 4,619 RMS intra-system matched records in the Analzyer Record Linkage application. The RMS fatal *source* dataset was matched into the HRVD fatal *target* dataset.[25, 26]

Phase 1 – Computer-generated matching

120. Strict matching (referred to as P1 matching) automatically identified "exact matches". Processing of "exact matches" via the automated P1 process eliminates the inefficiency of having a human compare every record in, or between the databases, with every other record.

121. Matching based on algorithms was applied to the data to generate a list of potential matches that were deemed to be highly probable. Calculations based on probabilities and frequencies of each data field within a record were weighted and ordered by rank, and a threshold level was established where the match being made was probably correct. The threshold was set after a review was made of the prospective algorithm-based matches, which eliminated the need for a human to compare every record for possible matches. Potential matches below that threshold were handled in one of two ways, depending on whether or not matching was for fatal or non-fatal, and intra- or inter-system matching.

122. For inter-system matching of fatal violations data the algorithm-generated match pools were imported into the Analyzer data matching system and the record linkage expert reviewed these computer-assisted match targets for each of the remaining unmatched source records. Non-fatal intra-system matching was completely automated with results reviewed by the record linkage expert to ensure that extreme over- or under-matching was not occurring.

Phase 2 - Computer-assisted matching

123. Computer-assisted matching, referred to as P2, was based on algorithms that generated pockets of potential matches between source and target records that were deemed to be likely matches, but required human review to select which of the closely weighted records was the best match. Calculations based on probabilities and frequencies of each data field between pairs of records were weighted and ordered by rank based on names, date of birth, date of death, place of birth and place of death. Using the Analyzer matching interface, the record linkage expert selected which target record from that pocket, if any, matched the source record being examined.

124. The P2 fatal inter-system matching rules were:

1. The sex of source and target(s) had to be equal, where sex was known.
2. The first initials of names between a source and target(s) had to be the same.
3. For target(s), where DOB and DOD were known, one of the dates had to be within 5 years of the source dates.

[25] An "exact match" occurs where two or more records in a database are matched together when all the fields on which matching decisions are being made are identical.

[26] The designation of source and target is determined by the number of records in the dataset. The smaller of the two datasets in the pair is the source and the larger is the target. This is to reduce the number or records that have to be compared, but each record from both datasets are compared to all of its potential matches.

4. If the source and potential targets(s) had “perfect” DOB or DOD, at least one of the other matching fields hand to match.

125. After the inter-system match work was done in Analzyer between the HRVD and RMS datasets, the resulting list of unique fatal victims was imported into a spreadsheet. The records were then sorted on the various data fields to determine if any other possible matches could be found. This not only served to catch matches missed, it also measured how good the matching algorithms had been. Additional fine tuning of algorithms was done as a result of the hand reviews by the record linkage expert, ensuring that successive matching passes would be more thorough and accurate.

Phase 3 – Vague data matching

126. In Phase 3 (P3) matching, records that contained too many blank fields, or were records of commonly-named individuals, from the same area, or who died in the same time period were matched. These matches did not have enough data to be specific about which source/target pair was exact, so one was randomly selected from the targets. For example, Mau Bere was a very common name in many parts of the country, and 1999 was a year when many of them died. It is unlikely that there were missed intra-system matches for two reasons. First, they were records that often came from the same statement which indicated they were family members with the same name. Second, the GCD recorded many deaths in the same cemetery with the same name and date (or no date), but there was not enough identifying information within the HRVD and RMS datasets to distinguish them as unique individuals.

127. The P3 matching process made matches where equal probabilities of a good match for a record existed, which did not require the judgement of the record linkage expert.

Pair-wise inter-system fatal matching

128. The inter-system matching pair of HRVD and RMS resulted in the new list of unique victims, named the HRVD/RMS dataset. This dataset included 10,594 records found only in the HRVD dataset, 4,087 found only in the RMS dataset, and 532 were found in both HRVD and RMS. These 15,213 total unique records were then inter-system matched with the 149,267 records of the GCD dataset, the HRVD/RMS dataset being the source data and the GCD the target data. The pair-wise matching between the HRVD/RMS dataset into the GCD resulted in 157,000 named deceased persons. This total includes records that were out of mandate or did not have dates of death to verify that they died within the mandate period. Only records having dates of death within the mandate period were used for analysis.

129. The linkages within and between these datasets are used in estimating the total number of dead due to the conflict. Records in this final list can linked back to a single dataset, or a combination of the three datasets. Below is a simple matrix showing the results of the final fatal inter-system matching linkages between the datasets.[27]

	HRVD only	RMS only	GCD only	HRVD & RMS	HRVD & GCD	RMS & GCD	HRVD/ RMS/GCD	Total
Count	5,203	2,148	141,787	382	5,391	1,939	150	157,000
Percent	3.31	1.37	90.31	0.24	3.43	1.24	0.1	100

130. If the intra-system matching caught all possible matches, then only zero or one potential match would have been possible during inter-system matching. Matches may be missed if the records being examined had missing data fields that made it unclear if the two records should have been linked. Human error is also possible when looking at the large quantity of data that was involved in the Commission’s work. Generally, a match is assumed when a majority of the data fields match, or the records' match weight is within tolerances. If

[27] These are unweighted totals, and they include records with missing dates, out of range dates, missing places, and places outside of Timor-Leste. Out of range records were subsequently eliminated from the analysis.

there are not enough fields with complete data, then it is difficult to determine with reasonable certainty whether a record should be included or excluded from matching to another. The latter case was especially true for the very common indigenous East Timorese names, like Mau Bere where many people, from the same place, died or were killed at the same time.

131. After completing the inter-system matching in Analzyer, the data were imported into a spreadsheet for review by the record linkage expert. By looking at the data sorted on different variables, with multiple processes—both human and automated—it can be confidently concluded that all possible matches that should have been made were processed. Additionally, the inter-system matching process may be considered a measure of Inter-Rater Reliability (IRR) because it finds instances where matches were missed in the intra-system phase. By returning to the intra-system data and applying the missed matches, it was possible not only to measure the IRR but also to correct the data, producing more reliable data on which estimates could be based.

Table 3 - Inter-system match record count totals & percentages for fatal violation by dataset pair

Step	HRVD to RMS	HRVD/RMS to GCD
Starting Count	HRVD + RMS=HRVD/RMS	
Spreadsheet Matching	Count & Percent	
Adjusted from Missed	Count & Percent	
HRVD/RMS total	Count & Percent	
Starting Count		HRVD/RMS + GCD = MSE
P1 Matching		Count & Percent
P2 Matching		Count & Percent
P3 Matching		Count & Percent
Total Count for MSE		Count & Percent

6. Data processing of reported violations involving groups of anonymous victims

132. During the statement-taking process a deponent may have talked about one or many victims. Sometimes when multiple victims were reported by a deponent, the deponent did not know some or all of the victims' names. In the Commission statement-taking process 1.9% (1,419/75,443) of victim-records which were documented by the Commission the deponent did not know the individual names of the victims, who suffered abuse as part of a larger group of people.

133. In order to integrate these data into Commission's analysis, and thereby consider violations against named individuals as well as unnamed groups, some further processing of the data was required to account for likely duplicate records of violations against a reported victim group. The processing steps to control for this duplication

- identified violation records (against unnamed group victims) which appeared to describe the same victim group, and then
- chose a victim record from the pool of possible duplicate records to be retained as the "rep rec" of this reported victim-violation.

134. Unlike data on violations against individuals (which by-and-large contain personal identifiers, such as names, ages and sex), violations reported against groups do not usually contain detailed identifiers of the victim-group. As a result, group-victim records were matched by comparing the following variables of each reported violation against a group:

- the district where the violation reportedly took place
- the violation-type into which the violation was coded, and
- the year and month in which the violation reportedly occurred.

135. Then after all the like group-victim records were matched together to form a cluster, the record with the largest group-size within each cluster was retained. All other records were regarded as duplicate records and therefore dropped from the dataset.

136. The level of duplication among group-victim records is shown in Figure <Duplication_ group_victim>. This table shows how many duplicate violation copies per violation type were identified in the dataset and the number of surplus group violation records which were dropped for the Commission's analysis on violations against group victims.

Figure <Duplication_group_victim>:

	Detention		Torture		Ill-Treatment		Displacement		Other Violations		All Violations	
Copies	Obs	Surplus	Obs	Surplus	Obs	Surplus	Obs	Surplus	Obs	Surplus	Obs	Surplus
1	441	0	134	0	121	0	180	0	736	0	1612	0
2	150	75	26	13	30	15	68	34	206	103	480	240
3	69	46	15	10	9	6	21	14	87	58	201	134
4	56	42	4	3	8	6	16	12	60	45	144	108
5	25	20	0	0	5	4	10	8	30	24	70	56
6	6	5	0	0	6	5	12	10	12	10	36	30
7	0	0	0	0	7	6	0	0	0	0	7	6
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	12	11	0	0	0	0	0	0	12	11	24	22
13	13	12	0	0	0	0	0	0	13	12	26	24
Total	772	211	179	26	186	42	307	78	1156	263	2600	620

7. Statistical estimation techniques used in the analysis of fatal violations and displacements

137. This section presents the survey-based estimation techniques and multiple systems estimation methods used to make the estimates of the total extent and pattern of mortality and displacement during the Commission's reference period.

RMS weight calculations

138. The survey sampling was described earlier: in 2003, the CAVR field teams interviewed 1,396 households selected from 138 aldeias and groups of aldeias, called clusters. The clusters were selected by a method called “Probability Proportional to Size” (PPS), and then ten (or 20) households were selected by simple random sampling in each cluster. If each cluster had exactly the same number of sampled households, the sampling probability of each household would be identical, a process known as “self-weighting”.[VIII] Due to sampling 20 households in multi-aldeia clusters and non-response in other clusters, not all clusters had the same number of sampled households; however, 78.5% of the sampled clusters have exactly 10 sampled households. Non-response was 3.1%, and so no non-response adjustment was made. The weights were calculated as follows.

139. For each cluster, the adjustment for varying cluster size is:

- cluster_adjustment = median_cluster_size / cluster_size

140. The raw 1990 household sampling probability is

- sp_1990 = (total number of sampled HHs) / (total HHs in 1990) = 1,396/168,858

141. And so, for each cluster, the pps weight is

- pps_wt_1990_raw = (1/sp_1990) * cluster_adjustment

142. There was considerable population change due to migration and growth between 1990 and 2004, when the survey was conducted. Before the weights could be estimated, the total number of households in each aldeia was adjusted from the 1990 census using data from the 2004 census. During the sample design, the clusters were chosen using the household counts for each aldeiea reported by the 1990 census. At the time these calculations were done (April 2005), the Census Timor-Leste 2004 enumeration data were available disaggregated only to the sub-district level, but not by suco (village) or aldeia.[28] Note that the 1990-2004 weight adjustments do not affect the total summed weight, which is fixed at the number of households that existed in 2004. The weight adjustments affect how much households in different places affect the projection.

143. Two sub-districts listed in the 1990 census were not listed in the 2004 census results: Fatumaca in Baucau was absorbed by Baucau sub-district, and in Oecusse, Pante Macassar B was subsumed in Pante Macassar. For these sub-districts, the number of households in 2004 was estimated by using the proportion of households in the absorbing and absorbed sub-districts in 1990 multiplied by the total in the absorbing sub-district in 2004.

144. Although the 2004 household totals are available from the census at the sub-district level, the RMS has too few responses at the sub-district level for the estimates of weights by sub-district to have adequate data (29 of the 59 sampled sub-districts have fewer than 20 responses). Therefore the 1990 weights were scaled to the 2004 district totals by the following calculation:

- district_adjustment = (Total HHs in 2004 in this district)/(Total 1990 weight in this district)pps_wt_2004 = pps_wt_1990_raw * district_adjustment

145. By forcing the weights to match the 2004 census's district household counts, the weights were normalised to sum to the total number of households in 2004 (194,943). The errors given in the results are calculated using Stata's standard survey modules.[IX] These modules use the survey design variables (stratum, primary sampling units and sampling weight) to make weighted estimates of the totals and Taylor-series approximations of the sampling errors. The error estimates assume random sampling with unequal sample weights. This assumption is conservative (ie. it will tend to underestimate the sampling error) with respect to weights calculated using the PPS methods described above [X] The data files used for these calculations are available at http://www.hrdag.org/timor.

RMS date assignment for displacement analysis

146. The survey asked respondents when they moved from each of their locations during the period 1974-1999. When respondents were uncertain of the specific date of their move, they often identified the year of the move and the point in the agricultural cycle or whether it was the dry or rainy season. For each of these partial or seasonal dates, we assigned the displacement to the quarter in which the period or season fell. Where the partial date identification could fall in more than one quarter, it was randomly assigned to a quarter. Of the 2,024 moves defined by the respondents as displacement events, 76.6% were identified at least to the quarter, and 15.7% more were identified by the season. Only 7.7% of the displacement events were identified by year without specifying the month.

[28] See http://dne.mopf.gov.tp for the census data.

RMS weight adjustments for mortality estimates

147. The calculation of the weights assumes that events reported by each household could only have been reported by that household. This assumption is the result of the weights being simply the reciprocal of the sampling probability for the given household. Therefore, if there were more than one household that could have given information about a specific death, the true sampling probability for that death is greater than the probability for a single household. Deaths reported by the survey respondents violate the single-reporting-household assumption because for each death, there may have been more than one household which could have given information about that death. Among the 5,402 total deaths reported by respondents, 545 were reported more than once (the duplicate reports were identified and removed before estimation). The duplicate reporting implicit in the survey weighting was corrected by adjusting the weights in the way described below.

148. Before the survey weights can be used to estimate the total number of deaths, they must be adjusted to account for the number of households that were potential respondents for each death. That is, for each death, how many relatives survived until 2003 to be potential respondents in the survey? Much of the information required for this calculation is available in the survey because the respondent's relatives are also the decedent's relatives. The number of surviving relatives for each decedent *D* was calculated based on the relatives reported by the respondent *R* using the following rules:

1. If *D* is a parent of *R*, the expected number of relatives surviving in 2003 is the sum of the following*:*

- Assume that *D*'s parents are 25 years older than *D* (or 50 years older than *R*, if *D*'s age is not reported); use age-specific conditional probabilities of survival (calculated from the survey) to estimate the expected number of parents alive in 2003
- Count *R*'s siblings as *D*'s children
- Given an average approximate total fertility rate of 5 prior to 1975, assume that *D* had four siblings with ages (-4, -2, +2, +4) years from *D*'s age (if *D*'s age missing, set *D*'s age to *R*'s age + 25), calculate the siblings' ages in 2003, and multiply each by the conditional probability of surviving to that age, and sum over four siblings.

2. If *D* is a sibling of *R*

- *D*'s parents are *R*'s parents, count the survivors directly
- *R*'s siblings are *D*'s siblings, count the survivors directly

149. Assume that D had the same number of surviving adult children as R.

3. If *D* is a child of *R*

- *R* and spouse are parents, count the survivors directly
- Adult children of *R* are *D*'s siblings, count the survivors directly
- Assume no surviving adult children of *D*.

150. This calculation yields the expected surviving adult relatives for each *D*, as well as indicating which of these surviving relatives live in *R*'s household, and which live in other households.

151. To convert the expected surviving adult relatives of *D* into an adjustment for the sampling weight, the number of relatives must be converted to an expected number of households in which the relatives live. There are on average 0.5 relatives of *D* (in addition to *R*) living in *R*'s household. Assume that other households in which *D*'s relatives live have the same concentration of relatives per household as *R*'s household (ie. 1.5 relatives per household). Thus, if *D* has *L* surviving relatives who live outside of *R*'s household, there are *a*

= 1 + *L*/1.5 households which could give information about *D*. The survey weights were adjusted for possible multiple reporting of *D* by dividing each *D*'s sampling weight by this factor, *a*. This calculation assumes that the other potential respondent households are in the same cluster as *R*, or that they are in a cluster with a similar within-cluster sampling probability.

Sensitivity analysis of assumptions in mortality re-weighting

152. There are a number of assumptions in the weight adjustments for the mortality estimates, including the following:

- The period difference between generations (assumed to be 25 years)
- The number of siblings respondents' parents had (assumed to be four)
- The birth spacing of parent's siblings (assumed to be two years)
- The number of adult children respondent's siblings had (assumed to be equal to the respondent's children).

153. These assumptions were tested using the following variations, and the annual total number of deaths were calculated:

- The inter-generational spacing was varied to 18 and 30 years
- The number of siblings respondents' parents were assumed to have was increased to six
- The birth spacing was increased to five years between siblings
- The number of adult children respondent's siblings had was assumed to be double the number of the respondent's children.

154. For each variant estimation, the annual totals were tested (by a two-mean t-test) against the main model. None of the years in any of the variant models was significantly different at $p<0.05$. The minimum p-value was 0.13, and it was an outlier: the second-lowest p-value was 0.23. Therefore, the estimates are not substantially sensitive to the assumptions about family structure.

155. Although the estimates are robust to the assumptions about family structure used to estimate the number of surviving relatives who could give information about D, the magnitudes of the estimates are sensitive to the model used to transform the estimated surviving relatives to estimated households that contain relatives. The estimated number of surviving relatives is L, and the estimated number of households containing relatives of a decedent D, denoted a, is a = 1 + L/1.5. The denominator 1.5 comes from the average number of relatives for D (including R) living in R's household (0.5). Varying this average from 0 to 3 (ie. assuming 1-4 surviving adult relatives per household) varies the resulting estimates of the total estimated deaths (by all causes) from -14.2% to +19.6%. The effect of varying this model declines over time, with the largest variations found in the early years 1972-1975 (-21%, +26%) and the smallest variations found in more recent years 2001-2003 (-11%, +16.2%). The decline is consistent over time.

156. Given a constant number of surviving relatives, fewer surviving relatives per household means more potential reporting households, a higher estimated sampling probability per reported death, and a lower sampling weight per reported death, and therefore fewer estimated total deaths; more adults per household reverses this logic.

157. Although the total estimates vary with changes in the model transforming relatives into households, the patterns are constant. The correlation coefficients for the main model to the low (0) and high (3) models above are each 0.99. Although the model of relatives-per-household does affect the total magnitude of the estimated deaths, it does not affect the estimated patterns over time.

Multiple Systems Estimation (MSE): motivation and theory

158. The survey analysis is conservative in the sense that it corrects for potential duplicate reporting by matching deaths across households, and because there is an adjustment to the sampling weights based on the estimated number of households which could have given information about each death. As some deaths may be reported by several households, there are other deaths which occurred during 1974-1999 for which there are no surviving relatives in 2003. If entire households died during the mandate period, there would have been no collineal relatives who could have given information in 2003. Given these limitations, an alternate method for estimating the total deaths may provide a check on the survey estimates.[29]

159. MSE uses several separately-collected incomplete lists of the population. The lists are matched identifying the elements common across lists in order to estimate the number of elements that are missing from all of the lists. In this project deaths documented in the HRVD, RMS, and GCD were matched across the three systems using the name, date of death, location of death and date of birth.

160. The most basic form of this technique is capture-tag-recapture, which uses only two lists.

161. A technical explanation of how a count of the unknown members of the population can be estimated is as follows. Consider the case of two projects P_1 (a list of *A* individuals) and P_2 (a list of *B* individuals). There are *M* individuals who are matched across both lists, in a universe of *N* total individuals (*N* is unknown). If all of the people in the universe *N* have an equal probability of appearing in List 1, then the probability of a specific individual being reported by P_1 is

162. *Pr(captured in list 1)* = $\frac{A}{B}$

163. Similarly, if all of the people in universe *N* have an equal probability of appearing in List 2, then the probability of a specific individual being reported by P_2 is

164. *Pr(captured in list 2)* = $\frac{B}{N}$

165. The probability of a specific individual being captured in both lists is

Pr(captured in list 1 and list 2) = $\frac{M}{N}$

166. By definition, the probability of an event composed of two independent events is the product of the independent probabilities. Therefore,

Pr(captured in lists 1 and 2) = Pr(captured in list 1) x Pr(captured in list 2)

167. Which is $\frac{M}{N} = \left(\frac{A}{N}\right)\left(\frac{B}{N}\right)$: given this equation, solve for *N*. Rearranging the terms, $\frac{M}{N} = \frac{(A)(B)}{N^2}$ and then multiplying by *N*, $M = \frac{(A)(B)}{N}$ multiplying again $(M)(N) = (A)(B)$, and

[29] This explanation follows P Ball, J Asher, D Sulmont, D Manrique, "How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000", a report to the Peruvian Truth and Reconciliation Commission. Washington, DC: AAAS. 28 August 2004. Available online at http://shr.aaas.org/hrdag/peru

finally dividing by M yields $N = \frac{(A)(B)}{M}$. Note that with the final equation, the total number of deaths *N* can be estimated using the totals from *A* and *B* and from the matches between them, *M*.

168. There are many assumptions implicit in this solution. For example, none of the lists has individuals reported twice and that matching between the lists is accurate. In this project these two assumptions were controlled during the data processing as described in the matching section.

169. Other assumptions inherent in the capture-tag-recapture model are more difficult to manage. First, the method assumes that individuals are not entering or leaving the universe during the process of creating the lists, and second that the lists were selected randomly from the population. In human rights documentation projects the first assumption is usually irrelevant because the documentation occurs retrospectively. The second assumption cannot be satisfied, and it must be replaced by the assumption that the estimation is robust to the selection process.

170. Another assumption is that the lists are independent, that is, that the probability that an individual is in list two is independent of the probability that the individual is captured in list one. The final assumption is homogeneity: that the individuals that compose the universe all have the same probability of being captured.

171. If either of these assumptions is violated, the capture-tag-recapture method will not yield an adequate estimate of the total population size. If there are more than two lists with adequate information, the problems of dependency or heterogeneity can often be managed through the specification and selection of appropriate models. However, in the data for the HRVD, RMS, and GCD, there are only two usable systems (RMS-GCD for deaths due to hunger and illness, and HRVD-GCD for killings).[30] Alone these estimates would be insufficient, but in combination with the RMS estimates, they provide useful additional information.

Allocating GCD by type of death

172. The graveyard data do not include the manner of death. There were 89,894 graves with at least a first initial (or name), a last name and a year of death between 1972 and 2003. Of these 7,117 matched either the HRVD or the RMS (or both), and through this match, the manner of death can be learned from the matched record's manner of death. The remaining 82,717 GCD records need to be allocated to the four categories of manner of death (killings, deaths due to hunger and illness, combatant deaths, and other deaths). From the RMS annual proportions of deaths by these four types are shown in Figure <number>, below. Note that these proportions exclude deaths for which the manner of death is unknown (204 of 3,235 deaths reported in the RMS between 1969 and 2004 have unknown manner of death).

[30] The initial application of multiple systems estimation to demographic estimation was by C Chandra Sekar and W Edwards Deming, "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, March 1949. A thorough discussion of the estimators for the dual-system approach and the relevant error calculations is available in Yvonne M M Bishop, Stephen E Fienberg and Paul H. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. 1975. For commentary on the use of these methods in human rights analysis, see Fritz Scheuren, "History Corner," *The American Statistician*, February 2004.

Table 4 - Figure <number>: estimated proportions of deaths, by period and manner of death

Period	Killing	Hunger/Illness	Combatant	Other
1972-1974	0.9%	95.9%	0.0%	3.2%
Margin of error	1.8%	5.1%	0.0%	4.9%
1975-1982	11.2%	83.0%	4.4%	1.4%
Margin of error	4.7%	5.1%	2.5%	0.6%
1983-1998	5.5%	86.5%	0.7%	7.2%
Margin of error	2.5%	3.7%	0.6%	2.5%
1999	16.2%	83.0%	0.4%	0.4%
Margin of error	10.2%	10.2%	0.8%	0.8%
2000-2003	3.5%	86.9%	0.8%	8.9%
Margin of error	3.1%	6.5%	1.6%	4.9%
Total	8.3%	85.1%	2.4%	4.3%
Margin of error	2.7%	3.1%	1.2%	1.2%

173. These proportions were used to allocate the unmatched GCD records to the distinct manners of death to be used in the MSE calculations for each year: the proportions from the period containing each year was used to allocate the GCD deaths in that year. The margin of error of the allocation was included in the estimated error for the MSE estimates.

Sensitivity analysis of the loss of social knowledge: adjustments for underestimates

174. The survey asked respondents about the deaths of their parents, siblings and children. However, some deaths left no parents, siblings or children still alive when the survey was conducted in 2004. If deaths occurred long in the past, even the decedents' children would have all died, leaving no one to report the deaths. In other cases small families may have suffered complete mortality, so that no one survived to report the deaths. As the survey estimates the number (or the rate) of deaths farther back in time, the underestimate resulting from the loss of social knowledge must become more severe. However, even in the nearly immediate past (for example, in 2003 for a survey conducted in 2004), it will be impossible to document some deaths which have left no survivors. For example, people who have no surviving parents, siblings or children who died in 2003 cannot be reported in the survey.

175. The crude death rate (per 1,000 people) is an estimate of how many people died, in total, by year. It is a standard demographic and health indicator, usually estimated by indirect methods using census records. For Timor-Leste, these rates are difficult to estimate because the quality of the 1980 and 1990 census data has been in dispute.[XI] The CDRs estimated by the US Bureau of the Census for Timor-Leste are shown for 1990-2004. The Indonesian overall rate is shown for 1983. The estimate shown for 1971 comes from an Indonesian government claim that in all of Indonesia between 1971 and 1990, the CDR declined by 45%; the 1971 estimate shown here is the 1990 estimate for Timor-Leste inflated by this factor. A projected CDR is also shown by linearly interpolating between the 1971 estimate and the 1990-2004 estimates.

176. In addition to the CDR estimates, the CDR from the Commission's RMS is shown. This estimate is the total estimated deaths divided by the estimated population for that year (multiplied by 1,000). There are several observations to be made about this graph. First, the CDR estimated by the US Census Bureau is within the confidence interval of the CDR estimated by the RMS beginning in 1993. In 2003 the confidence interval of the RMS CDR (4.2 – 6.6) contains the US Census Bureau estimate (6.4), as shown in the graph by the capped spike at the end of the CAVR line. That is, while the RMS greatly underestimates the death rate in the "normal" peacetime years 1972-1974, by the mid-1990s, the RMS agrees with the results obtained via the indirect methods employed by the US Census Bureau. This observation is consistent with the notion that the RMS estimates suffer increasing downward bias into the past.

177. During years in which the historical record suggests that substantial excess deaths occurred, the linear interpolation of the CDR underestimates deaths. These years include 1975-1979 and 1999.This is consistent with the literal meaning of "excess" deaths. (There are no census-based CDR estimates for the 1975-1979 period). Looking further into the past, the survey-based CDR captures a decreasing fraction of the total CDR (a similar graph can be drawn for the MSE estimates over time, with similar results).

178. To adjust the RMS, the deaths lost to the loss of social knowledge must be estimated over time. The model employed was the following:

- the number of deaths estimated by the CDR and the projected population for each year was estimated (CDR deaths), shown as a rate in Figure {g_cdrs.pdf};
- the fraction of CDR deaths that occurred due to hunger and illness was estimated using the fraction of all deaths reported in the survey that were due to hunger and illness (similar to the allocation used for the unmatched GCD data). In the survey the mean (and median) fraction of all deaths (over years) attributed to hunger and illness is 0.80, and 50% of all years are within the range 0.754 – 0.846;
- the ratio of estimated deaths to CDR_deaths was calculated for the peacetime years (1972-1974 and 2002-2003); this is the fraction of "rememberable deaths," called the "memory fraction;"
- The memory fraction for 1975-2001 was estimated by linear interpolation using the following equations:
- estimated memory fraction (MSE) = -39.1 + 0.0200*year
- estimated memory fraction (RMS) = -43.9 + 0.0224*year

- The memory fractions for MSE ranges from 0.241-0.936, whereas for the RMS, they ranged from 0.228 to 0.846. This difference has an enormous impact on the outcome.
- The adjusted estimate was calculated as the original estimate divided by the memory fraction for each year.

179. The adjusted estimates are presented below in Figures {g_huil_xs_mse.pdf} and {g_huil_xs_rms.pdf}. Note that in both graphs the raw estimates and the adjusted estimates converged as the year approached 2003. The impact of the higher memory fraction for the MSE relative to the RMS was apparent in the estimated total deaths in excess of the CDR baseline: the MSE adjusted estimate was 104,000 deaths while the RMS adjusted estimate was 183,300 deaths.

180. Both of these estimates depend on a number of assumptions, including assumptions about the shape of the decline of the CDR from the early 1970s through the late 1990s and about the nature of the loss of social memory. Smooth but non-linear changes in the loss of social memory (either concave up or concave down) would not change the estimate substantially. However, if the underestimates in the MSE and RMS due to social memory loss were somehow discontinuous or otherwise drastically different for 1972-1974 relative to the peak years 1975-1979, the adjustment employed here would not correct appropriately for the underestimate. Both of these models depend on CDRs calculated from the 1980 and 1990 census data and indirect methods used by the US Bureau of the Census. There is sampling and non-sampling error which is not represented in the graphs or the statistics, but the error is certainly substantial.

181. However, these models have the benefit of showing that with the adjustment, the estimated annual total deaths due to hunger and illness closely match the CDR baseline deaths for the pre-invasion period (1972-1974) and for the period 1984-1998.

182. There are several reasons to prefer the MSE estimate to the RMS estimate. Although the RMS more closely matches the CDR deaths estimate in the post-occupation years that approach peacetime, 2002-2003, the MSE more closely matches the pre-occupation CDR total deaths estimates. For the purposes of this estimate, the most relevant period is 1975-1979, and the choice of estimates should be guided by the best fit immediately before to this period. A second reason to prefer the MSE is that it is based on considerably more data than the RMS alone: the MSE uses the GCD data in addition to the RMS.

183. The strongest conclusion which can be made is that the unadjusted RMS and MSE estimates must be too low. Part 6: Profile of Human Rights Violations, provides an examination of statistical support for findings in relation to the number of fatal violations during the Commission's mandate period.

ENDNOTES

[I] UNTAET Regulation 2001/10 Section 13.1(a)(i).

[II] UNTAET Regulation 2001/10 Section 13.1(a)(i).

[III] UNTAET Regulation 2001/10 Section 13.1(a)(i).

[IV] UNTAET Regulation 2001/10 Section 13.1(a)(ii).

[V] UNTAET Regulation 2001/10 Section 13.1(a)(iv).

[VI] UNTAET Regulation 2001/10 Section 13.1(d).

[VII] Patrick Ball, *Who Did What to Whom Handbook*, and Patrick Ball et al, *HR Database Design Methods*. US.

[VIII] Paul S Levy and Stanley Lemeshow, Sampling of Populations, Chapter 11, Wiley, New York, 1999.

[IX] Stata Corporation, Stata Survey Data Reference Manual, v. 8, College Station, TX: Stata. 2003.

[X] Donna Brogan, "Sampling error estimation for survey data," in Household Sample Surveys in Developing and Transition Countries, United Nations Publication ST/ESA/STAT/SER.F/96, Department of Economic and Social Affairs of the United Nations Secretariat, 2005.

[XI] See for example, Ben Kiernan, "The Demography of Genocide in Southeast Asia: The Death Tolls in Cambodia, 1975-79, and Timor-Leste, 1975-80." *Critical Asian Studies* 35:4 (2003), pp. 585-597.