Anneks mengenai data dan metode statistik

1. Pendahuluan

1. Kejadian menyangkut hak asasi manusia adalah rumit. Seorang saksi mata atau korban bisa melaporkan tentang satu atau beberapa korban, yang masing-masing bisa saja mengalami satu atau banyak pelanggaran. Setiap pelanggaran bisa saja melibatkan satau atau banyak pelaku. Dengan demikian, interaksi antara orang-orang yang berbeda dalam ribuan kejadian sejenis ini mengharuskan adanya metode-metode identifikasi dan agregasi empiris yang cermat untuk mendukung analisis kuantitatif yang valid dan bisa dipercaya.

2. Untuk menjamin kualitas data, Komisi menempuh beberapa proses. Apendiks metodologis ini mengetengahkan data dan metode dari mana hasil statistik Komisi diperoleh.

3. Apendiks ini dibagi dalam enam bagian utama. Seksi 1 mengetengahkan garis besar tentang relevansi dari analisis data empiris terhadap mandat Komisi. Seksi 2 mengetengahkan gambaran terperinci tentang set-set data yang berbeda yang digunakan dalam analisis statistik Komisi. Seksi 3 menggambarkan editing data, pembersihan dan tehnik normalisasi nama yang diterapkan terhadap data. Seksi 4 mengetengahkan berbagai tabulasi perekaman pada tahapan yang berbeda dari proses konversi data. Seksi 5 mengetengahkan berbagai tehnik deduplikasi dan pertalian pencatatan yang digunakan untuk mencocokkan laporan-laporan berganda tentang korban perorangan yang sama. Seksi 6 mendokumentasikan proses pengolahan data yang digunakan untuk menerangkan laporan-laporan berganda atas kelompok-kelompok korban yang tidak dikenal. Akhirnya, seksi 7 mengetengahkan tehnik estimasi statistik yang digunakan untuk mendapatkan estimasi total dari magnitude dan pola dari pelanggaran dan pemindahan selama periode acuan Komisi.

Relevansi dari analisis data empiris terhadap mandate Komisi

4. Human Rights Data Analysis Group (HRDAG) membantu Komisi untuk mengumpulkan dan menganalisis data pelanggaran hak asasi manusia yang berhubungan dengan periode mandat Komisi, 1974-1999.[*] Apendiks ini menjelaskan bagaimana data ditata dan diproses.

5. Komisi mengharuskan sebuah system manajemen informasi untuk mengatur dan menyusun data yang dibutuhkan untuk menjawab persoalan yang digarisbawahi dalam mandatnya. Secara khusus, sistem manajemen informasi Komisi harus menyediakan informasi tentang pelanggaran-pelanggaran hak asasi manusia di masa lalu yang pada gilirannya harus menyediakan:

1. Analisis-analisis statistik deskriptif tentang pola-pola umum dan kecendrungan-kecendrungan pelanggaran supaya dapat menggambarkan “sifat” pelanggaran hak asasi manusia (jenis pelanggaran yang dilakukan).[1]
2. Proyeksi-proyeksi statistik dari total pelanggaran untuk menetapkan "tingkat" pelanggaran hak asasi manusia (jumlah total pelanggaran yang dilakukan).[2]

[*] HRDAG adalah sebuah divisi dari Benetech Inc di Palo Alto, California, Amerika Serikat. Para staf HRDAG termasuk para ahli statistik, programer komputer, dan ahli pencatatan. Para anggota tcatatanim HRDAG telah bekerja dalam proyek-proyek dokumentasi dan analisis hak asasi manusia berskala luas di lima benua, di lebih belasan Negara selama 20 tahun terakhir. HRDAG telah bekerja dengan komisi kebenaran yang resmi di Haiti, Afrika Selatan, Guatemala, Peru, Ghana dan Sierra Leone; dengan Pengadilan Kejahatan Internasional untuk bekas Yugoslavia; dan dengan kelompok-kelompok hak asasi non-pemerintah di El Salvador, Kamboja, Guatemala, Kolumbia, Afghanistan, Sri Lanka dan Iran. Untuk informasi selanjutnya lihat http://www.hrdag.org.

3. Hipotesis statistik yang menguji pengulangan dari pelanggaran-pelanggaran tertentu untuk menginvestigasi apakah pola-pola pelanggaran tertentu merupakan "sebuah pola pelanggaran yang sistematis".[3]
4. Analisis setingkat kasus dengan pengarsipan dan pencarian database untuk menjelaskan "anteseden-anteseden, keadaan, faktor-faktor, konteks, motif dan perspektif" yang mengakibatkan pelanggaran dalam skala besar.[4]
5. Analisis kuantitatif terstruktur dan test-test hipotesis untuk menyelidiki apakah "pelanggaran-pelanggaran hak asasi manusia merupakan hasil dari perencanaan yang disengaja, kebijakan atau otorisasi" dari pihak partai tertentu dalam konflik.[5]
6. Penjelasan resmi dari metodologi-metodologi keilmuan dan statistik yang digunakan untuk menunjukkan bahwa temuan-temuan Komisi didasarkan pada "informasi faktual dan obyektif dan bukti-bukti yang dikumpulkan atau diterima oleh Komisi atau diberikan sebagai bantuan".[6]

6. Komisi menyadari bahwa setelah mengalami pelanggaran-pelanggaran hak asasi manusia, banyak korban dan keluarga mereka hidup dalam kebisuan, ketakutan dan keterasingan, seringkali lebih dari 25 tahun lamanya. Karena itu Komisi harus memikirkan sistim-sistim manajemen pengumpulan data dan informasi yang dapat sekaligus menghasilkan data historis yang dapat dipercaya dan memajukan partisipasi public dalam proses-proses pencarian kebenaran.

2. Sumber-sumber Data

7. Seksi ini mengetengahkan ciri-ciri dari tiga database statistic utama yang dikembangkan oleh Komisi untuk melaksanakan analisis kuantitatif terhadap pelanggaran-pelanggaran hak asasi manusia di masa lalu dan memajukan rekonsiliasi di Timor-Leste. Human Rights Violations Database (HRVD) adalah sebuah kumpulan pernyataan naratif dari para korban, laporan kualitatif Amnesty International (AI) dan data-data yang dikumpulkan oleh Fokupers, sebuah organisasi non-pemerintah Timor-Leste. Retrospective Mortality Survey (RMS) adalah sebuah sample acak survey rumah tangga yang digunakan untuk memperkirakan pemindahan dan kematian selama periode mandat Komisi. Database Sensus Makam (Graveyard Census Database, GCD) adalah sebuah sensus menyeluruh tentang pemakaman umum di ke-13 distrik di Timor-Leste.

8. Paduan data dari ketiga aliran data Komisi — HRVD, RMS dan GCD—digunakan untuk membuat perkiraan demografis independent terhadap keseluruhan tingkat, pola, kecendrungan dan tingkat-tingkat pertanggungjawaban terhadap pelanggaran-pelanggaran fatal masa lalu di Timor-Leste.

Human Rights Violations Database (HRVD)

9. Seksi berikut ini akan menggambarkan tiga proyek dokumentasi yang dilakukan untuk membentuk Human Rights Violations Database dari Komisi. Proses transformasi informasi kualitatif dari proyek-proyek dokumentasi ini ke dalam data statistic juga diketengahkan. Akhirnya, rekaman perhitungan dari ketiga proyek dokumentasi juga diutarakan.

Proses pengambilan pernyataan oleh Komisi

10. Dalam bulan Februari 2003 Komisi mulai mengumpulkan pernyataan-pernyataan naratif dari orang-orang di ketiga-belas distrik di Timor-Leste dan dari orang-orang Timor-Leste yang saat itu tinggal di Timor Barat. Pernyataan-pernyataan ini adalah dasar dari HRVD. Komisi mengadakan kantor di ke-13 distrik untuk melaksanakan mandatnya. Setotal 7,669 pernyataan naratif yang relevan adalah dokumentasi terseleksi dari pelanggaran hak asasi manusia yang

dilaporkan. Naratif-naratif ini menyediakan informasi baik tentang pelanggaran fatal maupun pelanggaran non-fatal selama periode acuan Komisi.* Proses pengambilan pernyataan meliputi ke-65 subdistrik di keseluruhan 13 distrik di Timor-Leste.† Selain pengumpulan pernyataan di tingkat distrik, Komisi juga mengumpulkan 86 <s00120> pernyataan dari para pengungsi Timor-Leste dan orang-orang lainnya yang tinggal di Timor Barat, melalui kemitraan Komisi dengan sebuah koalisi dari LSM-LSM yang berbasis di Timor Barat.‡

11. Mengingat pemberian pernyataan secara keseluruhan bersifat sukarela dari pihak deponent, dan didasarkan pada sebuah sample yang mungkin, persebaran pernyataan di sepanjang lokasi-lokasi geografis tidaklah seragam. Sebagaimana ditunjukkan oleh grafik <g5000001>, Komisi pada dasarnya mengumpulkan lebih banyak pernyataan dari para deponent di Bobonaro dan Ermera daripada di distrik-distrik lain (lihat seksi dibawah untuk gambaran terperinci tentang faktor-faktor yang mungkin mempengaruhi proses pemilihan sample selama proses pengambilan pernyataan Komisi).

[INSERT <g5000001> about here]

12. Untuk menganalisis informasi kualitatif ini secara statistik, informasi-informasi tersebut di kodifikasi ke dalam sebuah database FoxPro dengan menggunakan standard-standar desain dari model data “Siapa Melakukan Apa Terhadap Siapa”.[7] Meskipun data-data ini menyediakan banyak pengetahuan yang bermanfaat, proses pengambilan pernyataan Komisi yang menghasilkan data-data ini tidak memakai sebuah sample acak berbasis kemungkinan. Komisi lebih cenderung menerima pernyataan dari mereka yang mau merelakan informasi yang bisa mereka ingat. Sebagai hasilnya, data naratif, secara tersendiri, tidak bisa dianggap sebagai yang secara statistik mewakili keseluruhan tingkat dan pola pelanggaran di Timor-Leste.

Ciri demografis dari para deponen

13. Sekitar 21.4% (1,642/7,669) <s00104> dari semua deponent dalam proses pengambilan pernyataan Komisi adalah perempuan. Dalam beberapa komunitas, para perempuan tidak berpartisipasi dalam kegiatan-kegiatan sosialisasi Komisi karena mereka diharapkan tinggal di rumah. Selain itu, hanya sedikit perempuan yang merupakan anggota dari organisasi-organisasi formal dengan akses kepada informasi menyangkut pekerjaan Komisi, dan sejumlah yang lain ragu atau malu untuk memberikan kesaksian.§

14. Komisi menerima pernyataan-pernyataan dari orang dewasa dari segala usia. Baik laki-laki maupun perempuan, jumlah tertinggi deponent adalah dari kelompok umur 40-44, sebagaimana terlihat dalam bagan <g500002>.

[Insert Figure <g500002.pdf> about here]

15. Meskipun ada perbedaan mendasar dalam tingkat partisipasi perempuan dan laki-laki dalam proses pengambilan pernyataan Komisi, deponen perempuan cenderung berbicara tentang pelanggaran terhadap diri mereka, (dalam hubungan dengan pelanggaran terhadap

---

* Tim-tim Komisi mengumpulkan sejumlah 7,824 pernyataan. Beberapa dari penyataan-pernyataan ini (155 pernyataan) tidak dimasukkan ke dalam HRVD karena tidak menyebutkan pelanggaran yang berhubungan dengan mandat Komisi ataupun pelanggaran yang disebutkan tidak dalam cakupan periode acuan Komisi.

† Tim-tim distrik dari Komisi secara umum bekerja dengan komunitas-komunitas sesuai dengan identifikasi secara local terhadap subdistrik-subdistrik, desa-desa dan aldeia-aldeia. Ketika Komisi memulai pekerjaannya pada awal 2002, jumlah dari subdistrik di Timor-Leste adalah 65; namun, Kantor Statistik Nasional dan Survey Suco Timor-Leste 2001 melaporkan 64 subdistrik.

‡ Koalisi LSM terdiri dari CIS (Center for Internally Displaced Persons Service), Truk-F, Lakmas (Lembaga Advokasi Kekerasan Masyarakat Sipil) Cendana Wangi, Yabiku and Yayasan Peduli |ndonesia (YPI). Staff dari LSM-LSM ini mengumpulkan pernyataan dari orang-orang Timor-Leste yang tinggal di Belu, Kefamenanu, Soe dan Kupang di Timor Barat antara Februari dan Agustus 2003.

§ CAVR, dokumen internal: Laporan Evaluasi Proses Pengambilan Pernyataan CAVR. Arsip CAVR.

orang lain) dalam proporsi yang kurang lebih sama dengan deponen laki-laki. Sebagai ditunjukkan dalam Figure <tDepSexVictSexM>, dari semua pelanggaran yang dilaporkan oleh perempuan, 30.6% (2,939/9,605) adalah pelanggaran terhadap diri mereka, sementara untuk deponen laki-laki, 35.3% (17,438/49,382) dari pelanggaran yang dilaporkan adalah terhadap diri mereka.

[Insert Figure <tDepSexVictSexM> about here]

16. Tantangan sosial, budaya dan ekonomi yang dihadapi perempuan dapat membatasi partisipasi mereka dalam sosialisasi dan proses pengambilan pernyataan Komisi. Namun, temuan-temuan statistik Komisi sesuai dengan klaim bahwa kebanyakan korban pembunuhan, penghilangan, penyiksaan dan penganiayaan adalah laki-laki muda. Sebaliknya, mayoritas terbesar pelanggaran seksual yang terdokumentasikan oleh Komisi dialami oleh korban perempuan (lihat Bagian 6: Profil Pelanggaran Hak Asasi Manusia).

17. Para pengambil pernyataan mewawancarai para deponent dalam bahasa Tetum, bahasa Indonesia atau bahasa atau dialek Timor-Leste lainnya (yang merupakan bahasa lisan meskipun tidak umum ditulis) dan kemudian menuliskan teks wawancara dalam bahasa Tetum atau Indonesia. Formulir pengambilan pernyataan tersedia dalam bahasa Tetum dan Indonesia. Dari 7,669 <s00101> pernyataan yang diterima Komisi dan ternyata berada dalam mandat Komisi, 81.7% adalah dalam bahasa Tetum, 17.0% dalam bahasa Indonesia, 1.2% dalam bahasa Timor-Leste lainnya, dan 0.1% dalam bahasa yang tidak ditentukan <s00002>. Karena formulir pengambilan pernyataan Komisi adalah dalam bahasa Tetum dan Indonesia, pernyataan yang diberikan dalam bahasa lain dituliskan oleh pengambil pernyataan ke dalam formulir resmi baik dalam bahasa Indonesia atau dalam bahasa Tetum sebelum kodifikasi, pemasukan data dan analisis pernyataan naratif.

Kemungkinan bias sample dalam proses pengambilan pernyataan

18. Sebagaimana dibahas dalam seksi dalam anneks ini sifat sukarela dari proses pengambilan pernyataan Komisi berujung pada sebuah tingkatan “seleksi sendiri”. “Seleksi sendiri” ini, pada gilirannya, menimbulkan sejumlah faktor yang mempengaruhi siapa saja yang bisa memberikan pernyataan seperti:

- Orang-orang yang tinggal di wilayah terpencil atau pegunungan yang sangat jauh dari tempat di mana data dikumpulkan (seperti kota kabupaten) mempunyai kesempatan yang lebih kecil untuk menjadi sampel dari pada mereka yang lebih dekat ke kota-kota setempat dan ibukota-ibukota distrik
- Orang-orang yang aktif secara sosial dan/atau tangkas secara fisik lebih mungkin untuk memberikan pernyataan dari pada mereka yang sakit, orang tua, penyandang cacat atau yang mempunyai trauma
- Orang-orang yang aktif dalam komunitas setempat atau berhubungan dekat dengan pejabat-pejabat desa-desa, subdistrik atau distrik setempat dan para tua-tua lebih besar kemungkinannya untuk berpartisipasi dalam proses sosialisasi dan pengambilan pernyataan karena usaha-usaha pengumpulan pernyataan secara lokal sering diorganisir lewat struktur dan pejabat desa setempat
- Orang-orang yang yang meninggal sebelum Komisi dibentuk tidak mempunyai kesempatan untuk mengungkapkan cerita mereka kepada Komisi; karena itu, peristiwa-peristiwa yang terjadi di masa lalu cenderung kurang sering dilaporkan dari pada peristiwa-peristiwa yang lebih kemudian
- Orang-orang dengan akses yang kecil atau tidak punya akses sama sekali kepada media dan komkhasasi massal lebih kecil kemungkinannya untuk mendekati Komisi, dan
- Orang-orang dari konstituen yang memusuhi Komisi lebih kecil kemungkinannya untuk membuat pernyataan.

19. Untuk menjawab masalah bias sampel, Komisi melengkapi proses pengambilan pernyataan dengan kumpulan pernyataan naratif dari Fokupers dan informasi dari sumber sekunder dari Amnesty International. Selain itu, untuk memperhitungkan bias dalam mengukur pemindahan dan pelanggaran fatal, Komisi mengembangkan Retrospective Mortality Survey yang mengumpulkan informasi terstruktur dari sebuah sample kemungkinan acak dari berbagai rumah tangga di Timor-Leste (Lihat seksi di bawah untuk pemaparan yang lebih lengkap dari desain tehnik sampling dan instrument survey yang digunakan untuk Retrospective Mortality Survey).

Amnesty International

20. Amnesty International melaporkan tentang situasi hak asasi manusia Timor-Leste selama periode mandat Komisi sebagian besar dengan cara mengumpulkan informasi melalui jaringan kerja bawah tanah di Timor-Leste dan melalui hubungannya dengan orang Timor-Leste diaspora di Australia dan Portugal.

21. Komisi menerima 322 laporan dan dokumen dari Amnesty International, yang disusun antara tahun 1975 dan 1999.[*]

22. Laporan kualitatif dan Urgent Actions dari Amnesty International dikodifikasi dan dimasukan ke dalam Human Rights Violations Database Komisi dengan menggunakan metode

[*] Komisi tidak berhasil mendapatkan Laporan-laporan Amnesty International berikut ini:
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
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
Sebagai akibatnya, analisis statistik Komisi terhadap pelanggaran-pelanggaran di Timor-Leste yang dilaporkan oleh Amnesty International tidak mencantumkan tindakan-tindakan dan insiden–insiden yang berhubungan yang tercakup dalam laporan-laporan ini.

dan standar yang sama yang digunakan untuk pernyataan yang dikumpulkan oleh Komisi. Informasi yang dikumpulkan dari Amnesty International menggambarkan situasi umum hak asasi manusia di Timor-Leste, karena hal ini dipantau oleh komunitas hak asasi manusia internasional pada saat itu.

Fokupers

23. Forum Komkhasasi Untuk Perempuan Loro Sae (Fokupers), sebuah LSM hak asasi manusia setempat, membangun sebuah database pelanggaran setelah kekerasan-kekerasan yang berhubungan dengan Konsultasi Rakyat di tahun 1999.[*] Database Fokupers database dibangun dari wawancara-wawancara berkelanjutan yang dilaksanakan oleh staf Fokupers dengan perempuan-perempuan Timor-Leste. Semula, tujuan utama dari wawancara-wawancara itu berkaitan dengan pekerjaan konseling yang dilakukan oleh Fokupers. Namun, tujuan-tujuan itu diperluas dengan memasukan dokumentasi untuk maksud-maksud penyelidikan oleh otoritas hukum yang berkompeten, seperti Unit Kejahatan Beban PBB. Pernyataan-pernyataan naratif diambil dalam bahasa Tetum.

24. Fokupers membangun databasenya untuk mendukung publikasi dari sebuah laporan tentang kekerasan terhadap perempuan. Database asli Fokupers dipusatkan pada mengetengahkan data biografis para korban, peristiwa-peristiwa naratif yang digambarkan, mengidentifikasi pelanggaran yang terjadi dan pelaku yang terlibat. Pada bulan Juli 2004, Fokupers menyerahkan data-data ini kepada Komisi dalam kesepakatan bahwa identifikasi perorangan dari pelaku, korban, atau anggota keluarga dalam database tidak akan diidentifikasi dalam Laporan Akhir Komisi. Staf Komisi mengkodifikasi kembali data-data tersebut, berdasarkan definisi-definisi terstandar dan skema penkodean Komisi, sehingga data-data ini dapat dianalisis sejalan dengan Human Rights Violations Database Komisi.

Penkodean sumber-sumber kualitatif (pernyataan naratif CAVR, Amnesty International dan Fokupers)

25. Pengkodean data adalah proses mentransformasi informasi naratif tak berstruktur tentang pelanggaran, korban, dan pelaku ke dalam sebuah satuan yang bisa dihitung dari elemen-elemen data, tanpa membuang informasi penting atau menyalahartikan informasi yang terkumpul itu.

26. Pada bulan Oktober 2003, tim pengolahan data Komisi memeriksa kembali proses pengkodean dan pemasukan data untuk mengidentifikasi kesalahan sistematis dan inkonsistensi dalam proses pengkodean dan pemasukan data. Pada saat itu, 2,473 pernyataan telah terkodifikasi dan dimasukan ke dalam database Komisi. Sebuah sample acak atas 15% dari pernyataan (yaitu 371 pernyataan) dalam database diambil, distratifikasi berdasarkan distrik di mana pernyataan tersebut diambil.

27. Setiap pernyataan diterima oleh seorang juru kode: sang pemberi kode mengkode kembali pernyataan itu tanpa melihat bagaimana pernyataan itu dikode sebelumnya. Lalu hasil kedua dibandingkan dengan pengkodean sebelumnya dan kesalahan-kesalahan dalam pengkodean pertama diidentifikasi, dicatat dan diperbaiki. Selain itu, juru kode juga memeriksa kembali entry database untuk pernyataan ini dan mengidentifikasi dan mencatat jika ada kesalahan dan memperbaikinya.

28. Dari 371 pernyataan yang diperiksa kembali, teridentifikasi 416 kesalahan pengkodean. 58% (241/416) dari kesalahan-kesalahan ini adalah kesalahan pengkodean pelanggaran, 12%

[*] Fokupers didirikan pada tahun 1997 untuk mendukung korban kekerasan politik melalui program konseling dan bentuk bantuan lainnya kepada perempuan korban pelanggaran, termasuk mantan tahanan politik, janda perang, dan istri para tahanan politik. Mandat Fokupers juga meliputi memajukan hak asasi perempuan di antara penduduk local, terutama perempuan Timor-Leste.

(49/416) kesalahan berhubungan dengan pengkodean afiliasi korban, 10% (42/416) berhubungan dengan tingkat kekhususan lokasi yang dikode 9% (36/416) berhubungan dengan afiliasi institusional dari pelaku. Dari 416 kesalahan pengkodean yang teridentifikasi, 70% (291/416) di antaranya adalah kesalahan tidak-diidentifikasi (yaitu, ketika tindakan tidak diidentifikasi sebagai sebuah pelanggaran atau orang atau lokasi tidak diidentifikasi oleh pemberi kode). 17% (71/416) kesalahan pengkodean karena pemberi kode menggolongkan sebuah tindakan sebagai sebuah pelanggaran meskipun apa yang digambarkan dalam naratif tidak memenuhi definisi dan kondisi-kondisi pembatasan dari kosa kata terkontrol dari Komisi. Akhirnya, 13% (54/416) kesalahan pengkodean adalah akibat dari uraian yang salah terhadap sebuah tindakan ke dalam kategori pelanggaran yang tidak benar.

29. Sebagai hasil dari pemeriksaan kembali terhadap pengkodean ini, tim pengolah data mengambil tiga langkah untuk mengurangi kesalahan-kesalahan ini selanjutnya: (1) sejumlah revisi dilakukan terhadap kosa kata terkontrol dari Komisi; (2) sebuah lokakarya pelatihan di mana hasil pemeriksaan kembali itu diketengahkan kepada tim pengkodean dan pelatihan lebih lanjut dilakukan dalam hal-hal tertentu yang diperlukan; dan (3) dilaksanakannya latihan penulisan kode dalam kelompok reguler di mana para pemberi kode mengkode pernyataan yang sama dan memeriksa kembali konsistensi dari pilihan pengkodean mereka dengan menggunakan review kualitatif maupun pengukuran Inter-Rater Reliability (IRR) kuantitatif.[*]

30. Jenis-jenis revisi utama yang dilakukan terhadap kosa kata terkontrol Komisi adalah:

- Pengurangan dalam kategori pelanggaran menjadi sebuah daftar yang lebih dapat diatur
- Penghalusan atas kondisi-kondisi yang membatasi dari kategori pelanggaran yang secara konseptual serupa (seperti penyiksaan dan penganiayaan)
- Memusatkan kembali kosa kata terkontrol hanya kepada pengukuran pelanggaran, tidak lagi sekaligus untuk pengukuran pelanggaran dan dampak fisik dan psikologis dari pelanggaran-pelanggaran ini
- Menyederhanakan definisi-definisi kategori pelanggaran dan menjamin sintaksis dari definisi tersebut lebih konsisten dengan kekhususan informasi yang dikumpulkan di dalam pernyataan-pernyataan (sebagai contoh, istilah-istilah tehnis hukum dibahasakan ke dalam bahasa umum atau dihilangkan, karena istilah-istilah itu tidak sesuai dengan kenyataan historis yang diukur)
- Revisi terhadap daftar aktor-aktor institusional; baik penyederhanaan daftar itu dan penyusunan hirarkis terhadap institusi-institusi untuk hubungan structural diantara institusi-institusi tersebut.

Hasil pengumpulan data HRVD

31. Kombinasi tiga sumber data HRVD menghasilkan sebuah database dengan pencatatan sebagaimana terlihat dibawah ini dalam Figure {ZZ}. Pencatatan ini mengetengahkan korban-korban perorangan dan kelompok, baik yang menderita pelanggaran fatal maupun pelanggaran yang non-fatal. Figure {ZZ} menunjukkan perincian jumlah pencatatan yang dikumpulkan dalam setiap database. Perhatikan bahwa angka-angka mewakili total data sebelum pembersihan dimana pencatatan yang invalid dan berganda dikeluarkan dari database.

Table 1 - Figure {ZZ}: Matriks pencatatan perhitungan untuk Database Pelanggaran Hak Asasi Manusia

	Jumlah Pernyataan	Jumlah individu	Pelanggaran Fatal	Pelanggaran Non-Fatal

[*] Inter-Rater Reliability adalah tingkatan di mana dua atau lebih pemberi kode setuju. Inter-Rater Reliability menjawab persoalan konsistensi dari implementasi dari sebuah sistem pengkodean.

	Jumlah Pernyataan	Jumlah individu	Pelanggaran Fatal	Pelanggaran Non-Fatal
Pernyataan-pernyataan CAVR				
Amnesty International				
Fokupers				
Total				

32. Kelompok adalah pencatatan tentang korban tak bernama yang mengidentifikasi dua atau lebih korban. Sejumlah korban mengalami pelanggaran non-fatal yang berganda, yang lainnya mengalami hanya satu pelanggaran fatal. Karena itu, total pelanggaran tidak sama dengan jumlah korban.

Retrospective Mortality Survey (RMS)

33. Komisi melakukan sebuah Retrospective Mortality Survey (RMS) untuk menyediakan sebuah perkiraan berbasis kemungkinan terhadap pemindahan dan kematian. Survei ini menarik sebuah sampel acak bertingkat dari rumah tangga, dan menggunakan angket terstruktur untuk mengumpulkan informasi tentng kematian dalam keluarga dan kejadian pemindahan selama periode acuan Komisi. Survey ini memungkinkan perkiraan statistik terhadap tingkatan kematian secara alami, kematian yang berhubungan dengn kelaparan, kematian yang berhubungan dengan konflik, dan perpindahan.

Sampel statistis yang digunakan dalam RMS

34. Sampel RMS didasarkan pada dua tahapan penyusunan sampel. Tahapan pertama adalah sebuah sampel dari 2,336 aldeia di Timor-Leste, dan tahapan kedua adalah sebuah sampel dari rumah tangga dalam aldeia-aldeia terpilih.[*]

35. Populasi dari rumah tangga ditingkatkan menurut variable-variabel berikut ini: kota/desa, lokasi distrik, dan tingginya populasi.[†] Metode stratifikasi implicit digunakan sehingga daftar aldeia disortir berdasarkan tingkatan variable-variabel berikut ini: kota/desa, distrik, dan ketinggian dari permukaan laut, dan sebuah sampel acak sistematis memilih aldeia-aldeia di setiap variable-variabel terstratifikasi.[‡] Sebuah pengukuran kumulatif dari besarnya variabel diciptakan dan sebuah sampel interval dikalkulasi sebagai jumlah kluster (144) dibagi dengan total pengukuran besarannya (180,015), yang setara dengan 1,250.1. Sekumpulan angka acak antara 1 and 1,250.1 diambil (397.235) dan aldeia dengan pengukuran kumulatif dari besaran di atas jumlah itu dipilih sebagai sampel. 1250.1 ditambahkan secara berulang ke angka awal dibangun secara acak dan aldeia yang diseleksi dari sepanjang daftar itu dalam cara yang sama.

36. Keputusan untuk menarik sebuah angka pasti dari 20 rumah tangga, ketimbang suatu angka yang proporsional terhadap besaran aldeia atau metode alokasi yang lain, terutama adalah suatu pertimbangan operasional. Menyeleksi sebuah jumlah pasti dari rumah tangga per

[*] Aldeia adalah unit administratif terkecil di Timor-Leste. Umumnya, sebuah aldeia adalah sebuah perkampungan dari sekelompok rumah di sebuah wilayah setempat. Biasanya, sebuah *suco* (desa) terdiri dari tiga atau empat aldeia, dan sekelompok *suco* membentuk satu subdistrik yang merupakan bagian administratif dari sebuah distrik. Menurut According Survey Suco Timor-Leste 2001 terdapat 13 distrik, 64 subdistrik, 498 *suco*, dan 2,336 aldeia di Timor-Leste. Tim distrik Komisi umumnya bekerja di seluruh 65 wilayah yang dianggap berbagai komunitas sebagai subdistrik, karena batas-batas administratif membutuhkan waktu untuk ditata kembali menyusul berakhirnya pendudukan.

[†] Stratifikasi adalah proses pengelompokkan anggota-anggota populasi ke dalam sub kelompok yang relatif homogen sebelum dilakukan sampling. Strata ini perlu terpisah (eksklusif) satu sama lain sehingga setiap unsur dalam populasi hanya dapat dimasukan ke dalam satu stratum. Strata tersebut juga harus lengkap secara kolektif, dimana tidak boleh ada unsur populasi yang diabaikan. Sampel acak dengan demikian diterapkan pada setiap stratum. Sampel acak bertingkat seringkali menambah baik keterwakilan sampel dengan mengurangi kesalahan penyamplingan.

[‡] Komisi menggunakan sebuah metode yang dikenal sebagai Kemungkinan Proporsional terhadap Besaran [Probability Proportional to Size] (dalam hal ini "size" menunjuk kepada jumlah rumah tangga dan bukan menunjuk kepada populasi, meskipun keduanya jelas saling berhubungan), sebuah rancangan umum dalam survey jenis ini.

aldeia adalah salah satu cara untuk menjaga kontrol atas keseluruhan besaran sampel dan agar mempunyai sebuah perkiraan penyebaran beban kerja yang sama di antara para pewawancara.

37. Komisi mempertimbangkan kelayakan menggabungkan responden Timor-Leste yang masih mengungsi di Timor Barat ke dalam populasi acuan.[*] Namun, keprihatian atas keamanan, operasional dan kualitas data yang muncul dari kondisi di Timor Barat membuat pelaksanaan survey di sana menjadi sulit. Karena itu, populasi acuan yang disampelkan oleh Komisi terdiri dari semua rumah tangga dalam ketiga-belas distrik di Timor-Leste.

38. Adalah tidak optimal, baik karena alasan-alasan statistik maupun operasional, untuk membolehkan aldeia-aldeia yang mempunyai kurang dari 20 rumah tangga untuk disampelkan. Karena itu, aldeia yang kecil digabungkan dengan aldeia di sekitarnya (yang tidak selalu harus berdekatan atau berbatasan), sebelum penyamplingan dilakukan, sehingga perkiraan jumlah rumah tangga dalam sebuah kluster (diartikan sebagai sebuah aldeia atau sekelompok aldeia) sekurang-kurang adalah 40, untuk mengurangi kemungkinan bahwa sebuah kluster sampel mempunyai kurang dari 20 rumah tangga. Dalam kenyataan, karena ketidakakuratan dari kerangka ini, setibanya di sebuah aldeia, sebuah tim lapangan bisa saja menemukan bahwa aldeia itu mempunyai kurang dari 20 rumah tangga, baik karena jumlah rumah tangga yang dilaporkan dalam sensus tahun 1990 tidak akurat, maupun karena telah terjadi perubahan dalam tahun-tahun antara. Karena alasan ini ke 144 kluster aldeia sampel sebenarnya terdiri dari 165 aldeia. Secara operasional, ini berarti bahwa dalam kluster-kluster ini, para pewawancara harus menarik sampel acak 20 rumah tangga dari antara gabungan jumlah total rumah tangga dalam kluster itu.

Desain Angket dan pengembangan Retrospective Mortality Survey

39. Angket RMS dirancang untuk memenuhi tujuan-tujuan berikut:

- Untuk menghasilkan perkiraan total kematian di Timor-Leste antara tahun 1974 dan 1999, dengan menggunakan baik tehnik-tehnik perkiraan berbasis survey dan tehnik-tehnik Perkiraan Sistem Berganda (Multiple Systems Estimation), dan
- Untuk mengembangkan analisis bebasis survey yang memperkirakan dan menggambarkan gerakan perpindahan yang rumit di Timor-Leste sepanjang periode mandat Komisi.

40. Sebagai hasilnya, angket itu ditata dalam modul-modul berikut ini:

- Sebuah register/daftar rumah tangga
- Sebuah daftar perpindahan kepala rumah tangga
- Sebuah riwayat kelahiran perempuan dewasa
- Sebuah riwayat saudara kandung laki-laki/perempuan dewasa
- Sebuah riwayat pengasuhan laki-laki/perempuan dewasa
- Sebuah seksi umum pelanggaran hak asasi manusia

41. Angket ini[†] direview oleh tiga ahli statistik hak asasi manusia yang bukan merupakan bagian dari Komisi[‡] dan beberapa orang spesialis di Komisi. Melalui proses review ini, perbaikan

[*] Seksi 3.3 Regulasi 2001/10 menyatakan: "Komisi dapat melaksanakan semua kegiatan yang konsisten dengan pemenuhan mandatnya dalam Regulasi saat ini."

[†] Lihat angket survey dalam Apendiks dari Annex ini.

[‡] Fritz Scheuren, Presiden dari Asosiasi Statistik Amerika (American Statistical Association), konsultan untuk HRDAG dalam proyek untuk Kosovo, Guatemala dan Peru; William Seltzer, Fordham University, dan Jana Asher, co-author laporan HRDAG di Kosovo, Sierra Leone dan Peru.

dilakukan terhadap tampilan dan rancangan angket, dan sejumlah masalah peristilahan dalam bahasa Indonesia dan bahasa Tetum diidentifikasi.

42. Satu paket yang terdiri dari delapan wawancara kognitif dilaksanakan selama tahap pengembangan angket itu. Pewawancaraan secara kognitif ini bertujuan menjelajahi proses kognitif dari responden. Pewawancaraan ini mencoba mengidentifikasi kesulitan dan kemungkinan jalan keluar terhadap tantangan-tantangan yang dihadapi oleh responden dalam (i) pemahaman terhadap pertanyaan, (ii) mengingat kembali informasi yang relevan, (iii) proses-proses keputusan, dan (iv) proses-proses tanggapan.[*] Sebanyak delapan orang —empat dalam kondisi kerja dan empat di lapangan— berpartisipasi dalam pewawancaraan secara kognitif ini. Pengertian yang berharga diperoleh dari penyelidikan terhadap ingatan responden terhadap tanggal. Secara khusus, proses dan tanggapan kognitif terhadap pertanyaan-pertanyaan yang berhubungan dengan waktu dan tanggal menunjukkan bahwa seringkali, ketika seorang responden menjawab “Tidak Tahu”, mereka mungkin hanya tidak mengetahui tanggal yang pasti menurut kalender Gregorian. Namun, tanggapan mereka menunjukkan bahwa seringkali waktu dari berbagai peristiwa lebih mudah diingat dengan merujuk kepada penanda waktu yang lain seperti peristiwa-peristiwa besar lainnya, atau titik tertentu dalam siklus pertanian atau musim.

43. Dari proses pewawancaraan kognitif, kami mengembangkan pemeriksaan tanggal terstruktur yang meminta responden untuk menyempitkan tanggal kejadian ke dalam sebuah “jendela enam-bulan” yang dapat ditentukan peristiwa-peristiwa besar seperti hari libur, atau petunjuk alam atau fisik (tingginya tanaman jagung atau tanaman lainnya, musim hujan atau musim kering). Proses pewawancaraan kognitif ini juga menunjukkan bahwa konsep waktu seperti “awal”, “pertengahan” dan “akhir” tidak dimengerti oleh semua responden, sehingga penyempitan jendela waktu lebih lanjut tidaklah mungkin.

44. Selama wawancara kognitif dan tes lapangan, responden sering hanya menjawab “Tidak Tahu” atau “ke gunung/hutan” sebagai tempat ke mana mereka mengungsi. Sebagai hasil pewawancaraan kognitif, satu seri pemerikaan yang teliti dibuat untuk mendapatkan gambaran yang lebih rinci tentang tempat di mana orang-orang mengungsi.

45. Setelah review diantara sesama staf Komisi dan proses pewawancaraan kognitif, angket final kemudian di terjemahkan dan diterjemahkan kembali ke dalam bahasa Indonesia dan Tetum. Angket tersebut kemudian diuji di lapangan selama 5 hari di aldeia-aldeia di Dili, yang bukan merupakan bagian dari sampel. Sebagai hasil tes lapangan ini, sejumlah kecil perbaikan pertanyaan berturutan, perbaikan gramatis, dan syntaksis dibuat.

Pelaksanaan survey dan pekerjaan lapangan

46. Dalam setiap rumah tangga sampel, kepala rumah tangga menanggapi baik registrasi rumah tangga (dalam mana semua penghuni rumah dicatat) maupun seksi perpindahan. seorang laki-laki dewasa kemudian secara acak menyeleksi dari populasi perempuan dewasa dari rumah tangga itu untuk menjawab modul riwayat kelahiran perempuan dewasa.

47. Sebelum meninggalkan setiap aldeia, semua angket diperiksa oleh supervisor lapangan untuk mengidentifikasi dan memperbaiki jika ada kesalahan dan inkonsistensi dalam angket yang sudah dilengkapi. Dua coordinator lapangan mendampingi tim yang terdiri dari 22 enumerator (penghitung) survey ke lapangan.

48. Dua belas aldeia yang dimasukkan dalam sampel tidak dapat dikunjungi oleh tim penghitung. Tim ini tidak dapat melakukan wawancara di ke-12 aldeia ini mengingat masalah keamanan pada saat itu. Figure {YY} mendaftarkan 12 aldeia yang tidak dihitung.

[*] Tourangeau 1984.

Distrik	Subdistrik	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. Selain itu, di beberapa aldeia kurang dari 10 rumah tangga yang dapat dihitung mengakibatkan sejumlah ketiadaan tanggapan (non-response) tambahan. Secara keseluruhan, dari 1,440 rumah tangga dalam kerangka sampel, terdapat 3.1% (44/1,440) angka non-response. Karena rendahnya angka non-response, tidak ada pertalian statistik nyata yang harus dilakukan untuk mengontrol non-response dalam survey ini.

Graveyard Census Database (GCD)

50. Untuk membangun data kematian dasar untuk Timor-Leste, Komisi melakukan sebuah sensus terhadap pemakaman umum di ke-13 distrik di Timor-Leste. Melalui proses ini informasi yang tersedia tentang nama, tanggal kelahiran, tanggal kematian, dan agama dikumpulkan. Batu nisan yang tidak mempunyai informasi sejenis ini juga dihitung dan ukurannya dicatat.[*] Dengan mengumpulkan informasi ini, Komisi menciptakan sebuah sistim registrasi vital *de facto* bagi populasi penduduk Timor-Leste. Yaitu, GCD membuat sebuah daftar dasar atas beberapa, atau mungkin hampir semua kematian, yang dapat digunakan untuk analisis kematian di luar proyek ini.

Pengumpulan data GCD

51. Untuk membantu sensus Komisi atas pemakaman umum di Negara ini, sebuah daftar dari pemakaman umum yang diketahui di Timor-Leste dilengkapi oelh staf lapangan CAVR atas konsultasi dengan para pejabat setingkat desa (suco), dan jika memungkinkan di tingkat aldeia. Sebuah "pemakaman umum" dalam penelitian ini diartikan sebagai sebuah lokasi yang disediakan secara khusus untuk pemakaman orang yang meninggal. Pengertian ini meliputi tempat pemakaman bersama yang berada di tanah milik umum atau tanah yang dimiliki oleh lembaga agama. Namun, tidak termasuk pemakaman keluarga yang bertempat di tanah milik pribadi.

52. Data GCD dikumpulkan oleh dua tim pengumpul data yang berbeda. Tim pertama mengumpulkan 128,751 catatan dari 803 kuburan, yang dimasukkan ke dalam serangkaian spreadsheet Excel. Tim pertama meliput bagian-bagian dari ke-13 distrik, namun hanya Dili yang diliput secara lengkap. Tim kedua menjangkau semua distrik, kecuali Dili, untuk menyelesaikan sensus ini. Mereka mengumpulkan 153,057 catatan tambahan dari 1,779 kuburan. Tim kedua menggunakan database FoxPro database untuk memasukan data mereka.

53. Tim pencatat Komisi mendokumentasikan semua batu nisan dalam pemakaman umum —baik yang ditandai maupun yang tidak ditandai. Kuburan yang bertanda adalah kuburan yang mempunyai struktur fisik yang mengenang kehidupan seseorang, dengan tulisan yang dapat

[*] Ukuran batu nisan tak mengandung informasi itu dapat digunakan sebagai sebuah indikator terdekat apakah orang yang meninggal itu adalah seorang anak-anak atau seorang dewasa.

dibaca dalam bahasa Inggris, Indonesia, Tetum maupun Portugis.[*] Pada semua batu nisan bertanda yang dapat dihitung, informasi berikut ini diberi kode jika terdapat dalam batu nisan: nama lengkap, tanggal lahir dan tanggal mati. Batu nisan tak bernama biasa berupa salib kecil sederhana atau tanda penguburan yang lainnya, tanpa nama atau informasi tentang tanggal kematian. Para penghitung diminta untuk mencatat informasi tentang agama, jenis bahan dan ukuran kuburan, jika informasi tersebut dapat dilihat dari batu nisan, baik untuk batu nisan yang bernama maupun yang tidak bernama.

3. Gambaran metodologis tentang tehnik-tehnik editing data, pembersihan dan normalisasi nama

54. Ketiga database yang digunakan oleh Komisi mengharuskan tehnik-tehnik editing data, pembersihan, dan normalisasi nama agar data-data tersebut bisa dibandingkan dan dikaitkan di antara database-database itu. Beberapa bulan dihabiskan untuk memeriksa kembali data-data ini atas kesalahan pengetikan atau pengejaan yang nyata, dan sebuah sampel acak pemeriksaan kembali dilakukan untuk menjamin akurasi data. Masalah tehnis muncul dalam pengalihan data dari struktur satu database ke database yang lain, dan ini juga diidentifikasi dan diperbaiki.

Pembersihan dan editing Database

55. Tim pengolahan data melakukan sebuah pemeriksaan lengkap (dan perbaikan di mana perlu) terhadap semua catatan HRVD menyangkut:

- Informasi yang hilang menyangkut distrik/subdistrik
- Informasi tanggal pelanggaran yang tidak masuk akal (misalnya hari = 42, bulan =13)
- Catatan-catatan dimana pelanggaran yang terjadi sebelum tanggal lahir korban
- Catatan-catatan dimana pelanggaran yang terjadi setelah tanggal mati korban
- Pernyataan dimana deponent dikode sebagai korban dari sebuah pelanggaran fatal
- Catatan-catatan dimana usia korban dikode sebagai 0 atau dengan angka negatif
- Catatan-catatan dimana usia korban dikode sebagai lebih dari 75 tahun
- Catatan-catatan dimana tidak ada kode pelanggaran yang tercatat
- Catatan-catatan dimana tidak ada korban yang tercatat untuk sebuah pelanngaran terkode
- Catatan-catatan dimana tidak ada pelaku (individual/institusional) ditetapkan dalam sebuah pelanggaran terkode.

56. Selain pemeriksaan kembali secara menyeluruh dan cepat sebagaimna digambarkan di atas, tim pengkodean juga memeriksa sebuah sampel acak sederhana dari catatan-catatan tentang pelanggaran fatal, penahanan, penyiksaan, penganiayaan, perekrutan paksa, pelanggaran berbasis seksual dan pemindahan. Maksud dari pemeriksaan secara cepat ini adalah untuk mengidentifikasi apakah ada kesalahan sistematis dalam afiliasi dari para korban dan tanggungjawab pelaku institusional. Satu inkonsistensi utama teridentifikasi – antara lain dimana afiliasi korban tidak diberikan untuk semua korban dari satu pelanggaran atau beberapa pelanggaran yang terjadi dalam tindakan yang sama atau tindakan-tindakan yang berkaitan erat dalam hal waktu. –Catatan-catatan ini diidentifikasi, dan aturan yang memadai diberlakukan untuk memberikan afiliasi korban secara benar di seluruh pelanggaran dalam tindakan yang sama atau aksi terdekat dari aktor yang sama.

[*] Karena kurangnya sumber, Komisi tidak dapat menghitung pemakaman Cina.

Editing dan pembersihan tanggal

57. Catatan yang jelas mempunyai kesalahan, seperti tanggal lahir, pelanggaran, atau kematian yang baru diakibatkan kemudian, diteliti dan dibetulkan. Hal ini terutama biasa terjadi dalam database GCD dimana penanda kuburan sangat kecil sehingga empat digit lengkap dari tahun tidak dapat dituliskan. Sistem pemasukan data tidak bisa menerima penanggalan tahun dengan dua digit, yang seharusnya seperti dalam 1900-an, sebagaimana juga dalam 2000-an. Para pencatat dari tim yang berbeda kadang menggunakan standard pengkodean tanggal yang berbeda. Ada yang menggunakan standar Eropa HH-BB-TTTT, ada yang menggunakan standar Amerika Serikat BB-HH-TTTT", ada yang menggunakan format TTTT-BB-HH, atau variasi dari standard-standar ini dengan menggunakan tahun dua digit. Lagipula, kadang tanda pemisah yang berbeda digunakan di antara tahun, bulan dan hari – antara lain "/", ".", dan "-". Sebagai akibatnya, semua format penanggalan di sepanjang ketiga dataset ini dipetakan ke dalam format terstandar, TTTTBBHH.

58. Jika Tanggal Lahir (TL) berada setelah Tanggal Mati (TM), tanggal-tanggal ini dipertukarkan. Dua jenis kesalahan yang menyebabkan penamaan bulan lebih dari 12 atau penamaan hari lebih dari 31 juga diidentifikasi dan diteliti. Komisi melihat bahwa beberapa kesalahan disebabkan oleh variasi setting format penanggalan pada komputer-komputer di mana data-data dimasukkan.

59. Kesalahan lain hanyalah kesalahan pengetikan. Catatan dari HRVD dan RMS diperbaiki dengan memeriksa kembali bahan-bahan catatan asli dan memasukkan perbaikan ke database. Untuk database GCD tidak cukup waktu untuk memeriksa kembali secara manual sumber-sumber itu, sehingga jika kesalahan tidak mudah dibetulkan, nilai dalam bidang penanggalan itu (bulan atau hari) dibiarkan kosong.

Editing dan pembersihan usia

60. Data usia diperiksa atas kemungkinan kesalahan pengetikan, sebgai contoh, orang-orang dengan usia di atas 100 tahun. Sumber dari pencatatan-pencatatan ini diperiksa kembali untuk memferifikasi data dan perbaikan dilakukan bilamana perlu. Jika Tanggal Lahir dan Tanggal Mati diketahui, usia ditetapkan. Nilai usia GCD dikalkulasi dan sebuah bidang penanggalan yang baru dibangun untuk membantu memudahkan pencocokan.

Editing dan pembersihan kode-kode pelanggaran dan hubungan

61. Pemeriksaan kembali dilakukan terhadap kode pelanggaran dan kode hubungan dalam HRVD dan kode RMS yang teridentifikasi tidak valid atau bertentangan dengan data lain dalam sebuah catatan tertentu (sebagai contoh, seorang perempuan dikode sebagai ayah). Kertas-kertas berkas-berkas sumber untuk pencatatan-pencatatan ini diperiksa kembali dan perbaikan dilakukan terhadap database.

Editing dan pembersihan kode lokasi geografis

62. Data lokasi geografis untuk database RMS dan HRVD dikode sesuai standard geokode Timor-Leste yang yang ditetapkan oleh pemerintah dan disetujui untuk digunakan oleh Komisi. Lokasi-lokasi dibagi dalam empat tingkat administratif —Distrik, Subdistrik, Suco (Desa), dan Aldeia. Untuk lokasi yang berada di luar Timor-Leste, kode untuk Timor Barat dan Jawa diciptakan dan jika lokasi tidak diketahui, lokasi-lokasi itu ditandai untuk kode yang terpisah untuk tempat yang tidak diketahui. Setiap pekuburan diberikan sebuah kode khusus yang disebut "id", untuk membedakan antara pekuburan-pekuburan dalam wilayah geografis yang sama.

63. GCD tidak dikumpulkan menurut standar kode geografis Timor-Leste, sehingga terjemahkan ke dalam kode-kode standar.

Deduplikasi pemakaman dan makam GCD

64. Ada beberapa hal yang menyebabkan terjadinya penggandaan pencatatan terhadap kuburan dan pekuburan dalam database.

- Tim-tim pengumpulan data yang berbeda secara kurang hati-hati melakukan pecatatan terhadap pekuburan yang sama. Banyak pekuburan tidak mempunyai papan nama penunjuk, menyebabkan sulit untuk mengidentifikasi catatan berganda hanya dengan nama mengacu kepada nama pekuburan.
- Lokasi suco (desa) dan aldeia tertentu seringkali sangat sulit ditentukan dalam beberapa wilayah pedalaman. Bahkan jika pekuburan mempunyai nama yang sama, bisa saja dikode ke dalam lokasi geografis yang berbeda. Tambahan lagi, banyak pekuburan mempunyai nama yang sama (Santa Cruz adalah nama yang paling umum), yang nama pekuburan saja tidak cukup untuk menentukan pekuburan yang digandakan dalam pengkodean untuk kode geografis yang berbeda.
- Banyak pekuburan di Timor-Leste tidak ditata secara linear. Hal ini kadang menyebabkan tim pencatat melewati batu nisan yang sama, dan mencatatnya lebih dari satu kali.
- Karena sangat banyaknya jumlah file-file kertas-kertas yang dibutuhkan untuk mengumpulkan semua data ini, ada kemungkinan duplikasi pemasukan data.

65. Adalah mungkin untuk menemukan keterkaitan antara id pekuburan dengan mengamati nama orang yang meninggal, lokasi pekuburan, nama pekuburan, dan tanggal lahir dan tanggal kematian yang lengkap setelah dicocokkan.[*] Ketika jejeran duplikasi ditemukan, salah satu pekuburan dikeluarkan dari dataset yang digunakan untuk analisis. Walaupun umum bagi banyak orang mempunyai nama depan dan nama belakang yang sama, dan kemungkinan kesamaan tanggal kematian, sangatlah tidak mungkin bahwa mereka mempunyai tanggal lahir dan tanggal mati yang sama. Oleh karena itu, Setiap pencatatan yang mengadung nama depan, nama belakang, tanggal lahir dan tanggal mati yang sama dianggap sebagai penggandaan, dan hanya satu pencatatan yang dipertahankan dalam database untuk analisis.

66. Tujuan dari proses deduplikasi GCD adalah untuk menjamin bahwa orang yang mati hanya dihitung sekali. Telah dipikirkan sejak awal bahwa selama pemindahan paksa orang-orang dapat semula dikuburkan di mana mereka meninggal, kemudian jenazah diambil oleh sanak keluarganya dan dikuburkan kembali di pekuburan di aldeia asal mereka. Juga dipikirkan bahwa jika jenazah tidak ditemukan, sebuah tanda peringatan di pekuburan setempat dapat saja dibangun atau pasang. Meskipun hal ini mungkin saja terjadi, pemeriksaan kembali secara teliti terhadap data tidak mengungkapkan bahwa penguburan kembali atau penandaan terhadap jenazah yang tidak diketemukan dengan sebuah batu peringatan merupakan praktek yang umum. Lagipula, ketika jenazah telah ditemukan, penanda pertama akan disingkirkan atau ditempatkan bersama-sama dengan jenazah, dengan demikian mencegah perhitungan berulang. Orang-orang mati yang tidak pernah dikuburkan atau yang tidak dikuburkan di pekuburan umum, tidak termasuk dalam GCD. Untuk dapat menghitung orang mati yang tidak termasuk dalam kesaksian HRVD, wawancara RMS, dan data kuburan GCD, kami melakukan perkiraan system berganda terhadap total kematian. Analisis ini dijelaskan dibawah.

[*] Sebuah pencatatan yang lengkap adalah pencatatan yang mempunyai hari, bulan, dan tahun baik untuk Tanggal Lahir maupun untuk Tanggal Mati. catatan

Proses-proses pembersihan nama-nama

67. Nama orang-orang dalam data Komisi perlu ditangani dalam dua cara. Pertama, nama-nama tersebut perlu diuraikan dalam tiga kategori nama —depan, tengah/kecil dan belakang. Setelah tahap ini selesai, kanonikalisasi nama diperlukan untuk membantu pertalian pencatatan. Kanonikalisasi adalah sebuah proses mereduksi setiap nama ke dalam kemungkinan bentuk yang paling sederhana dan paling signifikan, tanpa kehilangan keumumannya.

68. Nama-nama orang mengandung sejumlah variasi yang penting dalam pengejaan, dalam pembagian ke dalam tiga bidang nama, dan dalam tanda baca. Variasi nama-nama mempunyai banyak alasan. Dalam pernyataan-pernyataan naratif bersifat terbuka, seperti HRVD, deponen bisa saja merupakan keluarga dekat, teman, tetangga, atau kenalan jauh dari korban, dia bisa saja tahu atau tidak tahu bagaimana mengeja nama dari korban yang dilaporkan. Transkripsi oleh pengambil pernyataan dapat saja memberlakukan aturan ejaan dan tanda baca tambahan dan bahkan memasukan ejaan yang salah. Demikian juga perubahan ejaan dan tanda baca dapat terjadi pada tahap pengkodean data dan pemasukan data.

Penguraian nama

69. Untuk menjawab variasi penting dalam bagaimana nama-nama dibagi ke dalam tiga bidang nama; depan, belakang, tengah/kecil, nama-nama itu diuraikan menurut aturan yang ketat. HRDAG memutuskan untuk membagi nama-nama dengan menggunakan nama depan yang “pertama” untuk *depan*, dan nama belakang yang paling “belakang” sebagai *belakang*, dan dan semua nama yang lain dimasukkan ke dalam bidang *nama tengah/kecil*. Selain itu preposisi (contohnya, de, da, do, dos) dikeluarkan dari bidang nama karena penggunaan dari preposisi ini tidak konsisten dalam data.

70. Sebagai contoh, nama Portugis Maria Luisa da Costa da Silva dapat dimasukan ke dalam database sebagai:

Nama Depan	Nama Tengah/nama kecil	Nama Belakang
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. Proses penempatan nama akan menstandarisasi nama-nama ini sehingga sehingga nama depan adalah Maria sementara nama belakang hanyalah Silva. Semua nama yang lain, tanpa preposisi, dipindahkan ke bidang nama tengah/nama kecil.

72. Nama asli Timor-Leste Mau Bere dapat dimasukkan sebagai:

Nama depan	Nama tengah/nama kecil	Nama belakang
MAU BERE		
MAUBERE		
MAU		BERE
		MAUBERE

73. Penempatan nama dalam kasus ini akan menempatkan Mau di bidang nama depan dan Bere di bidang nama belakang.

Kanonikalisasi nama

74. Kanonikalisasi nama diterapkan kepada bidang nama depan dan nama belakang dari pencatatan setelah penempatan untuk membantu mencocokkan secara lebih mudah, khususnya alogaritma otomatis untuk hubungan pencatatan (record linkage). Variasi-variasi ejaan untuk nama-nama disaring ke dalam sebuah from representatif tunggal untuk setiap nama. Sebagai contoh, variasi ejaan berikut ini dikanonikalisasi menjadi AGUSTINO:

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

75. Nama-nama asli orang Timor-Leste lebih sulit untuk dikanonikalisasi karena umumnya hanya sepanjang empat atau lima huruf dan sejumlah pencatatan yang nampaknya merupakan variasi ejaan pada kenyataannya adalah nama yang jelas berbeda. Kanonikalisasi yang konservatif diberlakukan terhadap nama-nama asli orang Timor-Leste dan kemudian diuji dengan keterhubungan sampel dengan pencatatan animis yang melihat pada tanggal, usia dan tempat informasi untuk menentukan kanonikasi tambahan untuk diterapkan.

76. Setelah beberapa penempatan terhadap nama-nama tersebut untuk dikanonikalisasi, sebuah bidang baru dibuat dengan nama dieja secara terbalik. Kemudian, dengan mensortir bidang yang baru ini, kita dapat menemukan nama-nama tambahan untuk dikanonikalisasi menjado sebuah bentuk tunggal karena huruf awal dapat bervariasi menurut pengucapan, tetapi silabel akhir cenderung sama. Proses ini terbukti sangat menolong dalam menemukan kanonikal tambahan.

77. Ada juga nama-nama Cina, Indonesia (Muslim), dan nama-nama Anglo-Saxon di dalam database-database ini, seperti halnya nama-nama Portugis dan nama asli orang Timor. Relatif

sedikitnya jumlah nama-nama Cina, Indonesia dan Anglo-Saxon tidak mengharuskan penanganan khusus. Staf orang Timor-Leste, di Timor, mengidentifikasi apakah nama-nama adalah nama asli untuk penerapan aturan-aturan yang cocok dan alogaritma, karena nama-nama asli Timor-Leste tidak selalu spesifik menurut jenis kelamin.

78. Database-database HRVD dan RMS adalah lebih kecil dari GCD, sehingga kami mengkanonikalisasi database-database itu terlebih dahulu. Kemudian kami menerapkan daftar nama kanonikal terhadap GCD. Nama-nama hasil kanonikalisasi kemudian diperiksa kembali untuk mengidentifikasi kanonikal tambahan.

79. Selama proses kanonikalisasi, beberapa huruf dalam nama-nama ternyata bisa saling menggantikan satu sama lain, umumnya terjadi dengan nama-nama Portugis. Huruf-huruf S, J, G, dan Z sering saling menggantikan satu sama lain dalam nama-nama tersebut. Juga, huruf-huruf V, U, W, dan B juga sering saling menggantikan. Yang kurang sering terjadi adalah huruf-huruf H dan E saling menggantikan, atau dihilangkan begitu saja, contohnya Helder/Elder, Henrique/Enrique. Satu contoh dari saling menggantikan adalah nama Virginia, yang dapat dieja dengan B atau V. Sebagai contoh, variasi ejaan yang ditemukan untuk VIRGINIA kanonikal antara lain BIRGINIA, BERGINA.

80. Nama-nama yang dimulai dengan huruf-huruf ini dibandingkan satu sama lain untuk membantu dalam proses kanonikalisasi. Jika nama-nama itu mempunyai lebih dari satu huruf yang bisa saling menggantikan atau huruf yang bisa tergantikan itu berada di bagian tengah atau akhir sebuah nama, sangat sulit untuk menemukan kemungkinan kanonikal. Oleh karena itu, sebuah program ditulis yang membangun sebuah daftar nama di mana kombinasi dari huruf-huruf yang saling menggantikan cocok nama kanonikal yang lain. Ahli jaringan pencatatan mereview kombinasi-kombinasi ini untuk menentukan apakah kombinasi-kombinasi itu harus dikanonikalisasi atau kombinasi-kombinasi itu adalah nama-nama yang jelas khas. Jika ada kanonikalisasi tambahan karena adanya huruf yang saling menggantikan, huruf yang dicenderung dipilih untuk kanonikal adalah S (untuk S, J, G, dan Z), V (untuk V, U, W, B), dan H (untuk H dan E).

81. Sebagai tambahan, dalam proses kanonikal, diperhatikan bahwa ANJU dan ANJO sering dikutip sebagai nama depan atau satu-satunya nama dalam sebuah pencatatan. *Anju* umumnya digunakan untuk menunjuk kepada bayi yang meninggal dan sering ditemukan dalam pencatatan GCD jika seorang anak meninggal sebelum dibaptiskan dan karenanya tidak diberikan sebuah nama Kristen. Pencatatan dengan menggunakan *ANJU* dan sebuah nama belakang digunakan untuk proses pencocokan karena ada beberapa data yang teridentifikasi, namun pencatatan dengan *ANJU* semata sangat ambigu untuk dibuat keputusan yang beralasan untuk pencocokan.

Pengkodean jenis kelamin dan etnik

82. Selama proses kanonikalisasi, nama depan Portugis Portugis diperiksa kembali dengan frekwensi pengkodean jenis kelamin laki-laki, perempuan dan yang tidak diketahui.[*] Penkodean jenis kelamin yang jelas salah dibetulkan. Sebagaimana dengan kebanyakan nama- Latin, nama-nama yang berakhir dengan A umumnya seorang perempuan dan nama yang berakhir dengan O (atau U) biasanya laki-laki. Jika nama depan berakhir dengan huruf-huruf lain selain A, O atau U, frekwensi antara pengkodean laki-laki dan pengkodean perempuan diteliti dan jika perbedaannya besar, yang menunjukkan bahwa beberapa pencatatan telah dikode secara salah selama pemasukan data, maka perbaikan dilakukan terhadap database.

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

4. Konversi Data

83. Demi melancarkan semua langkah pengolahan data yang berhubungan dengan mencocokkan pencatatan yang berganda, setiap dataset ditransfer dari FoxPro aslinya atau program database Excel, ke program database Analzyer.[*] Skema FoxPro database pertama-tama digandakan dalam PostgreSQL untuk dmasukkan ke dalam Analzyer. Struktur database relasional untuk data HRVD dan RMS disimpan dalam Analzyer.

84. Figure {XX} menunjukkan jumlah total pencatatan dari setiap dataset yang dimasukkan ke dalam Analzyer. Harap dicatat bahwa jumlah total ini merefleksikan perubahan-perubahan pembersihan data yang mengakibatkan dibuangnya pencatatan-pencatatan yang berganda atau invalid.

Table 2 - Figure {XX}: Perhitungan total yang tercatat oleh pra dan pasca pembersihan database

Database	Pra-Pembersihan	Pasca Pembersihan	Perhitungan/% Fatal	Perhitungan/% Non-Fatal
HRVD	41456	37651		
RMS	4883	4619		
GCD				n/a [†]

5. Tinjauan Hubungan Pencatatan

85. Para individu yang dilaporkan dalam HRVD dan RMS kadangkala dilaporkan beberapa kali, oleh deponen yang berbeda dan mungkin juga muncul dalam catatan GCD. Untuk memastikan analisa statistik dapat mengontrol laporan-laporan berganda untuk orang yang sama, data tersebut membutuhkan hubungan pencatatan, juga dikenal sebagai pencocokan. Pencocokan diterapkan pada dua kategori umum untuk kajian ini – pelanggaran fatal dan non-fatal. Pelanggaran-pelanggaran fatal termasuk pembunuhan warga sipil, kematian karena penelantaran, penghilangan dan kematian penempur. Kategori pelanggaran non-fatal mencakup percobaan pembunuhan warga sipil, penahanan, penyiksaan, pemerkosaan, perbudakan seksual, kekerasan seksual, perlakuan buruk, pemindahan, perkawinan paksa, halangan atas hak-hak reproduktif, pengadilan tidak adil, penghacuran rumah-rumah, penghancuran ternak, pemerasan, pengancaman, perekrutan paksa and kerja paksa.

86. Ada dua jenis pencocokan yang dilakukan untuk keperluan perkiraan statistik; pencocokan intra-sistem dan antar-sistem. Pencocokan intra-sistem menghubungkan catatan-catatan yang mengidentifikasi orang yang sama dalam sebuah set data tunggal, dan masing-masing catatan bisa cocok dengan nol, satu atau banyak catatan lainnya dalam dataset tersebut. Pencocokan antar-system menggabungkan dua atau lebih daftar pencatatam yang khas dari sumber-sumber data yang berbeda secara bersamaan sehingga sebuah perkiraan sistem berganda dapat dilakukan. Pencocokan catatan-catatn dalam proses pencocokan antar-sistem hanya dapat cocok dengan nol atau satu cacatan lain dalam setiap dataset lainnya.

87. Karena rumitnya pencocokan antar- sistem dan pembatasan waktu kerja, data pelanggaran non-fatal dalam HRVD dan RMS hanya mempunyai pencocokan intra-sistem untuk statistik deskriptif. Data fatal, termasuk data GCD, telah dicocokan secara intra-sistem maupun antar-sistem sebagai landasan bagi kalkulasi perkiraan sistem berganda. Pencocokan dilakukan

[*] Analzyer adalah aplikasi sumber yang bebas dan terbuka yang digunakan untuk mengumpulkan, menyimpan dan menganalisis informasi tentang pelanggaran-pelanggaran hak asasi manusia berskala besar. Untuk informasi selengkapnya tentang Analzyer, lihat website HDRAG pada http://www.hrdag.org/resources/data_software.shtml.

[†] Bidang ini tidak berlaku untuk bidangGCD karena menurut definisi, seseorang yang dikubur di sebuah pemakaman berarti sudah meninggal.

menggunakan tiga metode: pencocokan manual dengan tangan, pencocokan oleh komputer dan pencocokan yang dibantu oleh komputer. Tiap metode ini dapat melibatkan lebih dari satu pass.*

Aturan-aturan Pencocokan

88. Tiap catatan tunggal dibandingkan ke semua catatan lainnya dalam setiap dataset untuk mencari pencocokan yang mungkin dan dianggap cocok ketika angka-angka penting dari nilai-nilai bidang cocok *persis*, *sangat dekat* atau *tidak bertentangan*. Bidang-bidang yang digunakan untuk pencocokan adalah: nama_depan, nama_belakang, usia, jenis kelamin, Tanggal lahir, Tanggal Mati, Tempat Lahir (place_of_birth, POB), dan Tempat Mati (place_of_death, POD). Bidang untuk nama tengah/nama kecil dan tempat wawancara juga digunakan untuk keperluan klarifikasi, tapi bukanlah merupakan bidang-bidang yang tersedia di ketiga dataset ini, dan seringkali jarang tersedia. Meskipun bukan merupakan bagian dari aturan-aturan pencocokan, data-data ini tetap dipertimbangkan oleh ahli perhubungan catatan. Tapi, tidak digunakan dalam pencocok otomatis oleh komputer manapun.

89. Keputusan-keputusan pencocokan yang digunakan untuk data Komisi cenderung mecocokan data secara berlebih.† Pencocokan berlebihan mengurangi jumlah catatan khas dan karena itu akan cenderung mengurangi perkiraan. Pencocokan berlebihan lebih sering dipakai dalam kasus-kasus dimana terdapat ketidakpastian bahwa suatu pencocokan akurat, untuk menghasilkan perkiraan-perkiraan yang konservatif.

Pencocokan nama-nama

90. Bidang-bidang nama depan dan nama belakang tidak selalu lengkap; sebagian hanya inisial atau tidak ada salah satu dari nama depan atau belakang. Diupayakan dilakukan pencocokan atas setiap catatan meskipun tidak lengkap, tapi untuk pencocokan fatal, catatan-catatan yang tidak memiliki nama depan atau belakang atau hanya berupa inisial, dihapus dari pencocokan karena tidak cukup data untuk membuat penilaian yang dapat diandalkan. Untuk pencocokan non-fatal, diupayakan untuk mencocokan pelanggaran-pelanggaran dengan DOB, DOD, dan informasi lokasi kematian dengan catatan-catatan lain dengan nilai sama pada bidang-bidang tersebut, meskipun tidak ada nama atau catatan tersebut hanya berupa inisial. Catatan-catatan dengan data nama yang kurang lengkap sangat bergantung pada tanggal dan tempat yang sempurna untuk dapat dicocokan dengan catatan-catatan lain. Banyak orang bisa saja meninggal pada hari dan tempat yang sama, dan untuk mengetahui orang yang mana untuk dicocokan dengan nama yang tidak lengkap, sangatlah sulit dan tidak dapat diandalkan.

Pencocokan jenis kelamin dan etnis

91. Bilamana jenis kelamin korban diketahui, hanya mungkin dapat dicocokan dengan catatan-catatan dari jenis kelamin sama atau yang jenis kelaminnya tidak diketahui. Catatan-catatan dimana jenis kelamin ditandai Tidak Diketahui bisa dicocokan dengan catatan-catatan yang dikodekan Laki atau Perempuan, tapi dalam grup yang cocok, kode-kode jenis kelamin tidak bisa bertentangan dengan catatan-catatan lain dalam grup tersebut.

* Pass adalah sebuah tinjuan atas semua data dalam sebuah dataset berdasarkan urutan atau alogaritma, untuk mencari kecocokan.

† Kelebihan pencocokan (over-matching) artinya hubungan-hubungan dibuat antara catatan-catatan yang mungkin saja tidak digandakan.

Pencocokan lokasi-lokasi

92. Kode-kode lokasi geografis yang digunakan untuk data Komisi dibagi dalam empat tingkatan: distrik, subdistrik, suco (desa), dan aldeia. Database GCD adalah satu-satunya dataset yang menguraikan informasi lokasi sampai tingkat aldeia, jadi tidak digunakan untuk keperluan pencocokan. Seringnya pemindahan-pemindahan membuat para saksi sulit untuk menunjuk secara tepat informasi lokasi, kecuali di tempat-tempat dimana pelanggaran terjadi, di tempat di mana saksi tinggal saat itu, atau tempat mereka semula mengalami pemindahan. Orang bisa saja mengalami pemindahan beberapa kali, di beberapa lokasi dan karena konflik berlangsung selama tiga dekade, untuk mengingat lokasi yang tepat bisa saja mengalami kesalahan-kesalahan.

93. Selain itu, batas-batas antara lokasi-lokasi geografis dipengaruhi oleh tiga faktor: perubahan nama tempat dan batas-batas geografis dari batas-batas adminitratif seiring waktu; tidak persisnya batas-batas, terutama di daerah pedalaman; dan kemungkinan kesalahan dalam pengumpulan, pengkodean dan pemasukan data. Karena itu, pencocokan-pencocokan di sana sini dalam suatu distrik dan di antara distrik-distrik yang berbatasan dipertimbangkan juga. Kemungkinan-kemungkinan kecocokan antara sebuah subdistrik dengan suco yang berdekatan juga diberikan preferensi yang lebih tinggi. Dalam mempelajari data secara teliti, catatan-catatan yang cocok pada lebih banyak bidang data selain tempat, memperkuat penilaian kami untuk pencocokan lokasi. Jika HRVD mendokumentasi sebuah kematian pada lokasi yang sama dengan lokasi wawancara, diasumsikan bahwa informasi lokasi sangat mungki akurat.

94. Dalam kasus-kasus yang langka, pencocokan dilakukan dengan melanggar aturan untuk data lokasi, tapi hanya jika jelas catatan-catatan tersebut menidentifikasi orang yang sama, dan bahwa kesalahan-kesalahan umum pengetikan diperhitungkan untuk perbedaan tersebut. Jika ada lebih dari satu kemungkinan pencocokan, algoritma pencocokan dicobakan untuk dicocokkan dengan catatan-catatan yang kurang spesifik agar dapat menyisakan catatan-catatan yang lebih spesifik untuk digunakan pada pencocokan berikutnya.[*] Jika ada persebaran yang merata antara lokasi dalam tingkatan geografis tertentu, lokasi yang kurang spesifik dipilih dan jika ada satu yang tidak lebih kurang spesifik, dari satu yang dipilih secara acak untuk menjadi "rep rec".[†]

Mencocokan tanggal

95. Karena konflik di Timor-Leste terjadi dalam waktu yang panjang, banyak responden tidak mengingat tanggal dan tempat pasti dari kejadian-kejadian yang berlangsung. Data GCD diasumsikan lebih akurat dalam informasi tanggal dan tempat karena mayat-mayat biasanya dikubur sesaat setelah kematian, dan dekat dengan tempat kematian. Ketika pencocokan dilakukan pada bidang tanggal, ahli hubungan catatan (record linkage expert) akan menghubungkan catatan-catatan yang lebih atau kurang tiga tahun dari yang lainnya. Jarang terdapat pengecualian untuk aturan ini, dan hanya dilakukan ketika bidang-bidang data lainnya cocok secara kuat. Catatan-catatan dengan data bulan dan tanggal sering tidak akurat dalam data HRVD dan RMS, karena ingatan cenderung salah dalam periode waktu lama seperti itu. Untuk itu, tanggal-tanggal yang lebih spesifik dicocokan satau sama lain jika tanggal-tanggal itu

---

* Catatan-catatan dihubungan bersama dan data yang khas untuk setiap catatan disimpan, karena yang bertentangan dengan duplikasi dihapus. Sebagai hasilnya, perlu dilihat keragaman dalam catatan-catatan yang cocok untuk mengetahui jika perbedaan-perbedaan itu akan merubah analisis secara mendasar.

† "rep rec" adalah catatan yang mewakili secara baik kelompok-kelompok catatan-catatan pencocokan dengan memiliki data yang paling lengkap. Catatan-catatan dengan tanggal dan tempat yang sangat umum dalam kelompok atau sebuah cacatan dengan tempat atau tanggal yang lebih pasti dianggap lebih lengkap. Semakin lengkap data, semakin baik setiap putaran pencocokan yang berikutnya dari pencocokan intra-sistem maupun pencocokan antar-sistem. Karena catatan-catatn ini dihubungkan bersama dan data yang khas untuk setiap catatan disimpan, dan yang merupakan duplikasi dihapuskan, sangat penting untuk memperhatikan variasi dalam catatan-catatan pencocokan untuk melihat jika perbedaan-perbedaan itu dapat mengubah analisis secara berarti.

berdekatan, dan dicocokan pada tanggal-tanggal yang kurang spesifik jika mereka tidak berdekatan.

Pembatasan tingkat pencatatan

96. Pembatasan pencocokan dilakukan untuk mencegah kelebihan pencocokan (over-matching). Secara khusus, pencocokan berikut tidak diperbolehkan:

- Catatan-catatan tentang korban-korban dari pernyataan yang sama (karena setiap pernyataan mengidentifikasi korban tertentu yang bisa saja mempunyai nama yang sama karena hubungan keluarga)
- Dua korban non-fatal tidak bisa dicocokan jika mereka dilaporkan dalam sumber catatan yang sama (karena pengkodean data dan metode-metode representasi database yang digunakan mencegah agar duplikasi catatan-catatan dari satu pernyataan tunggal dimasukkan dalam database).
- Seorang deponen tidak bisa dicocokan dengan sebuah pelanggaran fatal
- Sebuah catatan non-fatal tidak dapat dicocokan dengan sebuah catatan fatal jika tanggal tertentu yang dihubungkan dengan pelanggaran non-fatal terjadi sebelum Tanggal Lahir (DOB) catatan fatal
- Sebuah catatan non-fatal tidak dapat dicocokan dengan sebuah catatan fatal jika tanggal tertentu yang dihubungkan dengan pelanggaran non-fatal terjadi setelah Tanggal Mati (DOD) dari catatan fatal.

Pencocokan Intra-sistem

97. Dalam sebuah dataset, seseorang bisa saja diidentifikasi oleh beberapa saksi mata. Pencocokan intra-sistem menghubungkan catatan-catatan yang mengidentifikasi orang yang sama, untuk menghasilkan sebuah daftar orang-orang dengan nama khas, untuk mencegah kelebihan perhitungan, dan akhirnya, kelebihan perkiraan. Pencocokan intra-sistem sangat rumit dan sulit untuk dilakukan dalam sebuah database, karena seseorang bisa cocok dengan sejumlah *n* catatan lainnya dalam dataset. Karena itu, data diatur dalam sebuah spreadsheet, yang lebih memudahkan untuk mengatur dan mengatur kembali data-data tersebut dalam berbagai cara untuk menemukan hubungan-hubungan yang perlu dibuat.

98. Pencocokan intra-sistem yang dilakukan terhadap sebuah dataset sebelum catatan-catatannya digabung dengan dataset lain dapat memperlihatkan pola-pola yang melekat pada proyek pengumpulan data tersebut. Sebagian pola-pola ini mungkin saja kesalahan-kesalahan sistematis dalam pengumpulan, pengkodean atau pemasukkan data, atau bisa saja merupakan hasil dari struktur pengumpulan data tersebut. Pengamatan pola-pola ini dalam setiap dataset perlu dilakukannya, dan jika perlu, dilakukan perbaikan kesalahan-kesalahan yang digarisbawahi.

99. Ketiga dataset Komisi jika dikombinasikan akan menjadi sangat besar untuk memungkinan pencocokan data berkualitas tinggi, karena sejumlah pola itu tidak akan terlacak oleh mata manusia. Jika ketiga dataset itu dikombinasikan ke dalam sebuah daftar tunggal, hasilnya adalah daftar yang terdiri dari lebih 160.000 catatan. Mencari catatan-catatan yang cocok dalam daftar sepanjang ini akan sangat sulit bagi seorang pembaca.

Pencocokan fatal intra-sistim HRVD

100. Pertama-tama, pencocokan intra-sistem pada data fatal dalam HRVD dilakukan untuk menghubungkan catatan-catatan yang mendeskripsikan korban yang sama. Catatan-catatan itu dimasukkan ke dalam spreadsheet dan disortir berdasarkan nama depan, nama belakang, POD, dan DOD, untuk menemukan catatan-catatan yang cocok.

101. Saat catatan-catatan dihubungkan, sebuah “rep rec” dipilih.* Setelah setiap menyortiran, sebuah pass pencocokan dilakukan dan catatan-catatan yang berhubungan dalam sebuah grup pencocokan disembunyikan (tapi tidak dihilangkan) dari berkas data yang dihasilkan, sehingga hanya tinggal “rep rec”nya. Ini mengurangi *noise* dalam data-data tersebut. Noise diartikan sebagai catatan-catatan non- “rep rec” dalam sebuah kelompok pencocokan yang mengalihkan pencocok dari kemungkinan hubungan-hubungan “rep rec” dengan calon pencocokan lainnya. Semakin sedikit daftar catatan-catatan yang khas, semakin mudah untuk melihat kemungkinan kecocokan dan pola-pola lain di dalam data. Setiap pass berikutnya mengidentifikasi kecocokan tambahan, dan akhirnya, sebuah daftar catatan-catatan yang khas disaring dari keseluruhan dataset. Sekurang-kurang lima pass dilakukan pada tiap dataset.

102. Keseluruhan 15,043 catatan fatal dari dataset HRVD direduksi menjadi sebuah daftar yang terdiri dari 11,145 korban khas. Semua catatan itu kemudian dimasukkan kembali ke dalam sistem pencocokan data Analzyer. Catatan-catatan yang cocok dihubungkan kembli dengan “rep rec” untuk dianalisis jika semua pencocokan telah dilakukan.

Pencocokan fatal intra-sistem RMS

103. Pencocokan fatal intra-sistem RMS dilakukan dalam sebuah spreadsheet setelah pencocokan intra HRVD dilakukan. Pencocokan intra-sistem RMS menggunakan bidang-bidang yang sama dengan pencocokan intra-sistem HRVD dan juga melihat sumber dari pencatatan tersebut. Pencatatan-pencatatan kematian yang dikumpulkan dari rumahtangga yang sama tidak boleh dicocokkan satu sama lain karena pencatatan-pencatatan tersebut mengidentifikasi individu-individu khas, meskipun mereka memiliki nama dan DOD yang sama.

104. Sebanyak 4,883 catatan fatal dataset RMS dataset direduksi ke dalam sebuah daftar yang terdiri dari 4,619 korban khas.

105. Hubungan-hubungan yang dihasilkan dari dataset HRVD dan RMS dimasukkan kembali ke dalam model data Analyzer untuk digunakan dalam pencocokan yang dibantu komputer (computer-assisted) dan pencocokan yang dihasilkan komputer (computer-generated), dan untuk menghasilkan data untuk analisa. Informasi dan pola-pola yang didokumentasi oleh ahli hubungan catatan (record linkage expert) dalam tahap pencocokan menggunakan tangan kemudian digunakan untuk menghasilkan aturan-aturan pencocokan dan algoritma untuk proses-proses pencocokan yang dibantu komputer dan pencocokan yang dihasilkan komputer.

Pencocokan non-fatal intra-sistem HRVD

106. Alogaritma computer diatur untuk membersihkan dan mencocokan pelanggaran non-fatal dalam HRVD. Langkah ini disebut sebagai *auto-matching.* Algoritma pencocokan otomatis untuk pelanggaran-pelanggaran non-fatal dalam HRVD dikembangkan karena batasan waktu dan sumber daya tidak memungkinkan penggunaan seorang ahli hubungan catatan. Jumlah korban non-fatal tiga kali lipat dari jumlah korban fatal yang dilaporkan dalam HRVD.

107. HRVD mempunyai 41,546 catatan. Pencocokan otomatis intra-sistem menghasilkan sebauh daftar yang terdiri dari 37,651 korban khas pelanggaran fatal dan non-fatal.

Penilaian dan pencocokan auto-kanonikalisasi nama non-fatal

108. Langkah pertama dalam proses auto-kanonikalisasi adalah membuat sebuah tabel dengan versi yang sudah bersih dari semua (fatal dan non-fatal) nama asli dalam database. Untuk nama depan, versi-versi tersebut di *normalised*, *normalised-terse*, *first word of normalised*

* Sebuah “rep rec” adalah sebuah catatan yang mewakili sebuah grup yang dicocokan, karena ada lebih dari satu kemungkinan cocok.

(disebut first-namefirst), and *first word of normalised-terse* (disebut first-namefirst-terse). Metode yang sama diterapkan pada nama belakang, namun kata terakhir yang digunakan, bukan kata pertama. Lalu, untuk setiap nama korban dari pelanggaran non-fatal, dicoba untuk dicocokkan kombinasi-kombinasi nama lengkap non-fatal yang telah dinormalisasi berikut ini, pada semua nama lengkap fatal kanonikalis-tangan yang telah dinormalisasi:

- nama depan + nama bealakang
- nama depan -pendek + nama bealakang-pendek
- namadepan-pertama + namabealakang-terakhir
- namadepan-pertama-pendek + namabealakang-terakhir-pendek

109. Program pencocokan ini mencocokan satu set lengkap informasi sebelum mencoba mencocokan dengan informasi kurang lengkap. Pencocokan nama-nama non-fatal ke nama-nama fatal ini hanya dilakukan untuk nama-nama fatal yang telah dinormalisasi yang telah dipetakan ke dalam nama kanonikal yang khas; dengan informasi semakin pendek, jumlah nama yang dinormalisasi yang “diperbolehkan” untuk dicocokan makin sedikit (yang diimbangi dengan fakta bahwa lebih mudah untuk membuat kecocokan, karena informasi yang kurang pendek lebih tahan terhadap keragaman pengkodean dan kesalahan-kesalahan pemasukkan data).

110. Untuk nama-nama lengkap yang tidak bisa dikanonikalisasi, nama depan dan nama belakang dikanonikalisasi dengan independen. Urutan pencocokan nama depan adalah sebagai berikut:

- Nama depan
- Nama depan-pendek
- Nama depan-pertama
- Nama depan-pertama-pendek

111. Proses pencocokan berikutnya dikembangkan untuk mengikuti ronde pencocokan awal berdasarkan pada proses pembersihan dan pencocokan otomatis. Proses ini mencari kemungkinan kecocokan dengan nama-nama non-normalisasi dan mengidentifikasi kepadatan-informasi per bidang data dari setiap catatan nama. Persentase catatan-catatan yang mengandung nilai-nilai tidak kosong untuk masing-masing bidang data adalah sebagai berikut:

- 9% memiliki tanggal_lahir (semunya memiliki lahir_geo1)
- 44% memiliki lahir_suco_lokasi
- 50% memiliki lahir_subdistrik_lokasi
- 53% memiliki lahir_distrik_lokasi
- 70% memiliki Namadepan
- 94% memiliki Jenis Kelamin
- 100% memiliki Namabelakang (karena ini dalah bidang wajib yang dibutuhkan untuk pencocokan)

112. Karena bidang nama belakang adalah satu-satunya bidang non-kosong untuk semua catatan, hanya bidang ini yang bisa digunakan dalam index blocking. Blocking melihat catatan-catatan dimana bidang (-bidang) yang di-block memilikai nilai sama. Blocking untuk bidang nama belakang dilakukan pada empat huruf pertama tiap nama. Algoritma pencocokan harus dikalibrasi dengan hati-hati: jika terdapat banyak bidang kosong, maka kecocokan lebih dekat

dengan bidang-bidang non-kosong harus dilakukan (juga, kecocokan pada nama-nama belakang yang umum dihargai lebih sedikit).

113. Ada tiga jenis “kedekatan” yang bervariasi:

7. Jumlah huruf dalam nama yang cocok (4, 8, atau semua)
8. Jumlah tingkatan dalam lokasi kelahiran yang cocok (dari 1 sampai 3), dan
9. Kedekatan yang diperlukan dari tanggal-tanggal (dari 1/3 tahun sampai 3 tahun).

114. Dengan dua pertiga dari nama-nama korban telah dikanonikalisasi secara otomatis, dan satu set peraturan yang didefinisikan dengan baik untuk kedekatan pencocokan yang diwajibkan untuk jumlah berbeda dari bidang-bidang non-kosong, tingkat kecocokan yang dihasilkan adalah sekitar 15% (dibandingkan dengan 25% untuk data pelanggaran fatal yang dicocokan secara manual).

115. Tingkat kecocokan 15% untuk pelanggaran-pelanggaran non-fatal terlihat masuk akal karena:

- Hanya dua pertiga dari catatan-catatan nama yang dapat dikanonikalisasi, dan
- Biasanya diharapkan kepadatan laporan yang tinggi untuk pelanggaran-pelanggaran fatal karena biasanya pelanggaran-pelanggaran ini lebih mudah diidentifikasi dan lebih mudah diingat oleh banyak orang dalam jaringan sosial korban.

116. Sistem pencocokan otomatis inter-sistem pada non-fatal mengurangi dataset dari sejumlah 44,203 catatan menjadi sebuah daftar dengan jumlah 31,568 catatan korban khas.

Review pencocokan non-fatal intra-sistem HRVD oleh ahli hubungan data

117. Ahli Penghubungan Data mempelajari sampel hasil-hasil pencocokan-otomatis untuk memastikan tidak ada kesalahan pencocokan (yaitu kelebihan pencocokan). Tidak ada pola sistematis kelebihan pencocokan yang ditemukan dalam peninjauan kembali sampel acak jika 10% dari grup yang cocok dicatat. Grup catatan terbesar yang dicocokan satu sama lain adalah 20 catatan. Peninjauan dilakukan pada grup terbesar untuk memastikan kalau ukuran kecocokan mereka dapat terima.

118. Pencocokan intra-sistem untuk data fatal menghasilkan daftar kombinasi dari individu-individu khas yang semuanya telah meninggal, meskipun penyebab kematian berbeda-beda. Ketika pencocokan intra dilakukan pada pelanggaran-pelanggaran non-fatal, seorang korban dapat mengalami satu atau lebih pelanggaran, dalam satu atau lebih hari, pada satu atau lebih tempat. Pencocokan non-fatal memperlihatkan pelanggaran-pelanggaran HAM yang dialami korban-korban individual, dimana seorang korban mungkin mengalami pelanggaran-pelanggaran lain yang mungkin atau mungkin saja tidak menimbulkan kematian.

Pencocokan Antar-sistem

119. Pencocokan antar-sistem menghubungkan daftar-daftar individu yang khas dari berbagai dataset dan dilakukan secara kumulatif dalam pasangan-pasangan atau dataset-dataset. Pencocokan antar-sistem hanya diterapkan pada data fatal. Pertama, pencocokan antar-sistem diterapkan dengan menggunakan 11.126 catatan cocok intra-sistem dari HRVD pada 4.619

catatan cocok intra-sistem RMS dalam aplikasi Analyzer Record Linkage. Dataset *sumber* fatal RMS dicocokan dengan dataset *target* fatal HRVD. [*]

Tahap 1 – Pencocokan yang Dihasilkan Komputer

120. Pencocokan ketat (disebut sebagai pencocokan P1) secara otomatis mengidentifikasi "pasangan pasti".[†] Dengan menghasilkan yang "pasangan pasti" melalui proses P1 otomatis menghilangkan inefisiensi yang ada jika manusia yang harus melakukan perbandingan setiap catatan di dalam, atau di antara database, dengan setiap catatan lainnya.

121. Pencocokan berdasarkan algoritma diterapkan pada data untuk menghasilkan sebuah daftar pasangan potensial yang dianggap tinggi kemungkinannya. Kalkulkasi yang didasarkan pada probabilitas dan frekuensi tiap bidang data dalam sebuah catatan dipertimbangkan dan diurutkan berdasar ranking, dan tingkat ambang batas ditentukan dimana pencocokan yang sedang dibuat kemungkinan benar. Ambang batas ditentukan setelah dilakukan peninjauan dari prospektif pencocokan berbasis algoritma, yang mengeliminasi kebutuhan akan tenaga manusia untuk membandingkan setiap catatan untuk pasangan yang mungkin. Pencocokan potensial di bawah ambang batas tersebut ditangani dalam satu dari dua cara, tergantung apakah pencocokan dilakukan untuk fatal atau non-fatal, dan pencocokan intra-sistem atau antar-sistem.

122. Untuk pencocokan antar-sistem untuk data pelanggaran-pelanggaran fatal, kumpulan pencocokan yang dihasilkan algoritma dimasukkan ke dalam sistem pencocokan data Analyzer dan seorang ahli hubungan catatan (linkage expert) mereview target-target pencocokan yang dibantu komputer tersebut untuk setiap sumber catatan yang tidak tercocokan. Pencocokan non-fatal intra-sistem sepenuhnya otomatis dengan hasil yang ditinjau oleh ahli hubungan catatan untuk memastikan kelebihan atau kekurangan ekstrim pencocokan tidak terjadi.

Tahap 2 – Pencocokan dengan bantuan Komputer

123. Pencocokan dengan bantuan komputer, disebut dengan P2, dilakukan berdasar algoritma yang menghasilkan banyak kecocokan potensial antara catatan-catatan sumber dan target, yang dianggap sangat mungkin, tapi memerlukan peninjauan oleh manusia untuk memilih mana dari catatan-catatan yang tertimbang yang merupakan pasangan yang paling baik. Kalkulasi berdasar probabilitas dan frekuensi tiap bidang data antara pasangan-pasangan catatan dipertimbangkan dan diurutkan berdasar nama, tanggal lahir, tanggal kematian, tempat lahir dan tempat kematian. Menggunakan interface pencocokan Analyzer, ahli hubungan catatan memilih catatan target mana dari kumpulan tersebut, jika ada, yang cocok dengan catatan sumber yang sedang diperiksa.

124. Aturan-aturan pencocokan intersistem fatal P2 adalah:

1. Jenis kelamin sumber dan target (-target) harus sama, dimana jenis kelamin diketahui.
2. Inisial pertama dari nama-nama antara sumber dan tujuan (-tujuan) harus sama.
3. Untuk target (-target), di mana DOB and DOD diketahui, salah satu tanggal harus berada dalam jarak 5 tahun dari tanggal sumber.
4. Jika sumber dan target(-target) potensial mempunyai DOB atau DOD yang "sempurna", paling tidak salah satu dari bidang pencocokan lainnya harus sama.

---

[*] Penunjukkan sumber dan target ditentukan oleh jumlah catatan dalam dataset. Dari kedua dataset, yang lebih kecil dari kedua dataset menjadi sumber dan yang lebih besar adalah target. Ini untuk mengurangi angka atau catatan yang harus dibandingkan, tetapi tiap catatan dari kedua dataset dibandingkan terhadap semua pasangan potensialnya.

† Sebuah "pasangan pasti" terjadi ketika dua atau lebih catatan dalam sebuah database cocok satau sama lain ketika semua bidangbidang dimana keputusan pencocokan dibuat, identik.

125. Setelah pencocokan antar-sistem dilakukan antara dataset HRVD dan RMS dalam Analyzer, daftar para korban fatal khas yang dihasilkan dimasukkan ke dalam sebuah spreadsheet. Catatan-catatan itu kemudian diurutkan pada berbagai bidang data untuk menentukan apakah kecocokan yang mungkin lainnya dapat ditemukan. Ini tidak saja berfungsi untuk mendapatkan kecocokan yang terlewat, tapi juga mengukur sebaik apa fungsi algoritma pencocokan tersebut. Penyempurnaan algoritma tambahan dilakukan sebagai hasil dari tinjauan ahli hubungan catatan, untuk memastikan pass-pass pencocokan suksesif akan lebih teliti dan akurat.

Tahap 3 – Pencocokan data yang tidak jelas

126. Dalam pencocokan Tahap 3 (P3), catatan-catatan yang mengandung terlalu banyak bidang kosong, atau catatan-catatan individu bernama umum dari area sama, atau yang meninggal dalam periode waktu sama, dicocokan. Pencocokan ini tidak memiliki cukup data untuk secara spesifik tentang pasangan sumber/target mana yang tepat, jadi satu dipilih secara acak dari tujuan-tujuan tersebut. Sebagai contoh, Mau Bere adalah nama yang sangat umum di berbagai tempat di negeri ini, dan 1999 adalah tahun dimana mereka meninggal. Kemungkinannya kecil ada pencocokan intra-sistem yang terlewat, karena dua alasan. Pertama, mereka adalah catatan-catatan yang sering datang dari pernyataan kesaksian yang sama, yang mengindikasikan mereka adalah anggota keluarga yang memiliki nama sama. Kedua, GCD mencatat banyak kematian dalam pemakaman yang sama dengan nama dan tanggal sama (atau tanpa tanggal), tapi tidak banyak informasi yang dapat menidentifikasi dalam dataset HRVD dan RMS untuk membedakan mereka menjadi individu-individu khas.

127. Proses pencocokan P3 menjadikan kecocokan-kecocokan dimana ada probabilitas yang sama akan kecocokan bagus sebuah catatan, yang tidak memerlukan penilaian ahli hubungan catatan.

Pencocokan fatal pasangan-baik antar-sistem

128. Pasangan pencocokan antar-sistem HRVD dan RMS menghasilkan daftar baru korban-korban khas, yang disebut dataset HRVD/RMS. Dataset ini termasuk 10.594 catatan yang ditemukan hanya dalam dataset HRVD, 4.087 catatan yang ditemukan hanya dalam RMS dataset, dan 532 catatan yang ditemukan dalam HRVD dan RMS. 15.213 catatan-catatan khas total ini kemudian dicocokan secara antar-sistem dengan 149.267 catatan dalam dataset GCD. Dataset HRVD/RMS menjadi sumber datanya dan GCD target datanya. Pencocokan secara berpasangan antara dataset HRVD/RMS ke dalam GCD menghasilkan 157.000 nama orang yang telah meninggal. Total ini termasuk catatan-catatan di luar mandat atau tidak memiliki tanggal kematian, yang bisa memverifikasi apakah mereja meninggal dalam periode mandat. Hanya catatan-catatan yang memiliki tanggal kematian dalam periode mandat yang digunakan untuk analisa.

129. Hubungan-hubungan dalam dan antara dataset digunakan dalam mengestimasi jumlah total kematian karena konflik. Catatan-catatan dalam daftar akhir ini dapat dihubungkan kembali ke dalam dataset tunggal, atau kombinasi dari ketiga dataset. Di bawah ini adalah matriks sederhana yang menunjukkan hasil dari pencocokan fatal antar-sistem terakhir yang mencocokan hubungan antara dataset-dataset.*

	HRVD saja	RMS saja	GCD saja	HRVD & RMS	HRVD & GCD	RMS & GCD	HRVD/ RMS/GCD	Total
Jumlah	5,203	2,148	141,787	382	5,391	1,939	150	157,000

* Ini adalah total yang tidak dipertimbangkan, termasuk catatan-catatan tanpa tanggal, tanggal di luar jangka waktu, tanpa tempat, dan tempat di luar Timor-Leste. Catatan-catatan di luar jangkauan ini kemudian dihilangkan dari analisa.

	HRVD saja	RMS saja	GCD saja	HRVD & RMS	HRVD & GCD	RMS & GCD	HRVD/ RMS/GCD	Total
Persentasi	3.31	1.37	90.31	0.24	3.43	1.24	0.1	100

130. Jika pencocokan intra-sistem menangkap semua kecocokan yang mungkin, maka hanya nol atau satu kecocokan potensial yang mungkin selama pencocokan antar-sistem. Kecocokan bisa terlewat jika catatan-catatan yang sedang diperiksa memiliki bidang-bidang data yang hilang, yang menjadikannya tidak jelas apakah kedua catatan tersebut harus dihubungkan. Kesalahan manusia juga mungkin ketika melihat besarnya kuantitas data yang ada dalam pekerjaan Komisi. Secara umum, sebuah kecocokan diasumsikan ketika mayoritas bidang data cocok, atau bobot kecocokan catatan-catatan tersebut masih dalam toleransi. Jika tidak terdapat cukup bidang dengan data lengkap, sulit untuk menentukan dengan kepastian yang beralasan apakah suatu catatan seharusnya disertakan atau tidak disertakan dalam pencocokan dengan catatan lainnya. Kasus yang ke dua terutama benar untuk nama-nama asli Timor yang sangat umum, seperti Mau Bere, dimana banyak orang, dari tempat sama, meninggal atau dibunuh pada waktu yang sama.

131. Setelah menyelesaikan pencocokan antar-sistem dalam Analyzer, data tersebut kemudian dimasukkan ke dalam sebuah spreadsheet untuk ditinjau oleh ahli hubungan catatan. Dengan melihat data yang diurut berdasar variabel-variabel berbeda, dengan proses berganda – baik oleh manusia atau otomatis – dapat secara yakin disimpulkan bahwa semua kemungkinan pencocokan yang seharusnya dilakukan, telah diproses. Selain itu, proses pencocokan antar-sistem dapat dianggap sebagai suatu pengukuran Inter-Rater Reliability (IRR), karena menemukan contoh-contoh dimana pencocokan terlewati dalam tahap intra-sistem. Dengan kembali pada data intra-sistem dan menerapkan pencocokan yang terlewat, memungkinkan untuk mengukur bukan saja IRR tapi juga memperbaiki data, yang menghasilkan data yang lebih bisa diandalkan yang dengannya perkiraan dapat dilakukan.

Table 3 - Total catatan perhitungan pencocokan antar-sistem & persentasi pelanggaran fatal menurut pasangan dataset

Langkah	HRVD ke RMS	HRVD/RMS ke GCD
Mulai perhitungan	HRVD + RMS=HRVD/RMS	
Pencocokan Spreadsheet	Hitung & Persen	
Diperbaiki dari yang terlewat	Hitung & Persen	
HRVD/RMS total	Hitung & Persen	
Mulai perhitungan		HRVD/RMS + GCD = MSE
Pencocokan P1		Hitung & Persen
Pencocokan P2		Hitung & Persen
Pencocokan P3		Hitung & Persen
Total Perhitungan untuk MSE		Hitung & Persen

6. Pengolahan data pelanggaran-pelanggaran yang dilaporkan melibatkan kelompok-kelompok korban yang tidak dikenal

132. Selama proses pengambilan pernyataan, seorang deponen mungkin saja telah berbicara tentang satu atau banyak korban. Terkadang, ketika korban berganda dilaporkan oleh seorang deponen, ia tidak mengetahui beberapa atau semua dari nama-nama para korban tersebut. Di dalam proses pengambilan pernyataan Komisi, 1.9% (1,419/75,443) dari catatan tentang korban yang didokumentasikan oleh Komisi, deponen tidak mengetahui nama korban-korban yang mengalami pelanggaran sebagai anggota dari kelompok yang lebih besar.

133. Untuk mengintegrasi data ini ke dalam analisa Komisi, dan maka menimbang pelanggaran-pelanggaran yang dilakukan terhadap individu-individu yang diketahui namanya maupun kelompok-kelompok yang tidak diketahui namanya, beberapa pengolahan data lebih lanjut dibutuhkan untuk memperhitungkan kemungkinan duplikasi pencatatan pelanggaran terhadap kelompok korban yang dilaporkan. Langkah-langkah pengolahan untuk mencegah duplikasi ini

- mengidentifikasi catatan-catatan pelanggaran (terhadap kelompok korban tak dikenal) yang nampak digambarkan sebagai kelompok korban yang sama, dan kemudian
- memilih sebuah catatan korban dari kelompok catatan-catatan kemungkinan duplikasi untuk dipertahankan sebagai “rep rec” dari pelanggaran yang dilaporkan tersebut.

134. Tidak seperti data pelanggaran terhadap individu (yang sedikit tidaknya mengandung pengidentifikasian pribadi, seperti nama, umur dan jenis kelamin), pelanggaran-pelanggaran yang dilaporkan atas kelompok pada umumnya tidak mengandung pengidentifikasian pribadi yang terinci dari kelompok korban. Akibatnya, catatan-catatan korban-kelompok dicocokkan dengan cara membandingkan variable-variabel berikut dari setiap pelanggaran yang dilaporkan terhadap sebuah kelompok:

- distrik di mana pelanggaran dilaporkan terjadi
- jenis pelanggaran yang ke dalamnya pelanggaran tersebut dikodekan, dan
- tahun dan bulan pada waktu pelanggaran dilaporkan terjadi.

135. Kemudian, setelah semua catatan-catatan jenis korban-kelompok dicocokkan bersama untuk membentuk sebuah kluster (kelompok sampel), catatan dengan jumlah kelompok paling besar di dalam setiap kluster dipertahankan. Semua catatan yang lain dianggap sebagai catatan duplikasi dan karenanya dikeluarkan dari dataset.

136. Tingkat duplikasi di antara korban-kelompok ditunjukkan di dalam Figur {X}. Tabel ini menunjukkan berapa banyak salinan pelanggaran duplikasi per jenis pelanggaran yang diidentifikasi dalam set data dan jumlah catatan pelanggaran kelompok surplus yang dikeluarkan dari analisa Komisi tentang pelanggaran atas korban-kelompok.

	Penahanan		Penyiksaan		Penganiayaan		Pemindahan		Pelanggaran lain		Semua pelanggaran	
Salinan	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. Tehnik-tehnik perkiraan statistic yang digunakan dalam analisis pelanggaran fatal dan pemindahan

137. Bagian ini menjelaskan tentang teknik estimasi berbasis survei dan metode estimasi system berganda yang digunakan untuk membuat estimasi dari total tingkatan dan pola kematian dan pemindahan selama masa acuan Komisi.

Kalkulasi beban RMS

138. Sampling survei telah dijelaskan sebelumnya: pada tahun 2003, tim-tim kerja lapangan CAVR mewawancarai 1,396 rumah tangga yang dipilih dari 138 aldeia dan kelompok-kelompok aldeia, yang disebut sebagai kluster. Kluster-kluster ini dipilih dengan sebuah metode yang disebut sebagai "Probability Proportional to Size" (PPS) dan kemudian sepuluh (atau 20) rumah tangga dipilih dengan sampel acak sederhana dari dalam setiap kluster. Apabila setiap kluster memiliki jumlah rumah tangga sampel yang persis sama, probabilitas sampel dari setiap rumah tangga akan sama, sebuah proses yang disebut sebagai "self-weighting".[8] Karena sampling atas 20 rumah tangga dalam kluster multi-aldeia dan non-respon dalam kluster lainnya, tidak semua kluster memiliki jumlah sample rumah tangga yang sama; namun, 78.5% dari kluster sample memiliki tepatnya 10 rumah tangga yang dijadikan sample. Non-respon adalah 3.1%, sehingga tidak ada penyesuaian non-response yang perlu dilakukan. Beban statistik kemudian dihitung dengan cara berikut ini.

139. Untuk setiap kluster, penyesuaian untuk ukuran kluster yang berbeda adalah:

- kluster_penyesuaian = median_kluster_ukuran / kluster_ukuran

140. Kemungkinan kasar sampling rumah tangga 1990 adalah

- sp_1990 = (jumlah total Rumah Tangga (RT) tersampel) / (total RT di tahun 1990) = 1,396/168,858

141. Maka, untuk setiap kluster, beban PPS adalah

- pps_beban_1990_kasar = (1/sp_1990) * kluster_penyesuaian

142. Ada perubahan populasi yang cukup besar karena migrasi dan pertumbuhan antara 1990 dan 2004, ketika survei ini dilakukan. Sebelum beban statistika ini bisa diperkiraan, jumlah total rumah tangga di setiap aldeia disesuaikan dari sensus 1990 dengan menggunakan data dari sensus 2004. Di dalam disain sample, kluster-kluster dipilih dengan menggunakan jumlah rumah tangga untuk setiap aldeia yang dilaporkan pada sensus 1990. Pada waktu kalkulasi ini selesai (April 2005), data enumerasi sensus Timor-Leste pada tahun 2004 hanya tersedia terpisah-pisah berdasarkan tingkat sub-distrik, tapi tidak berdasarkan suco (desa) atau aldeia.[*] Patut dicatat bahwa penyesuaian beban statistika 1990-2004 tidaklah mempengaruhi jumlah keseluruhan beban, yang dipatok pada jumlah rumah tangga yang ada pada tahun 2004. Penyesuaian beban ini mempengaruhi sebanyak apa rumahtangga-rumahtangga di tempat-tempat yang berbeda mempengaruhi proyeksi data.

143. Dua subdistrik yang terdaftar pada tahun 1990 tidaklah terdaftar pada hasil sensus tahun 2004: Fatumaca di Baucau menjadi bagian dari subdistrik Baucau, dan di Oecusse, Pante Macassar B dimasukkan ke dalam Pante Macassar. Untuk kedua subdistrik ini, jumlah rumahtangga pada tahun 2004 diperkirakan dengan menggunakan proporsi rumah tangga di

* Lihat http://dne.mopf.gov.tp untuk data sensus.

dalam wilayah subdistrik yang menyerap dan terserap pada 1990 dikali dengan total rumah tangga di subdistrik yang menyerap di tahun 2004.

144. Walaupun jumlah total rumah tangga 2004 tersedia dari sensus pada tingkat subdistrik, RMS memiliki terlalu sedikit respon pada tingkat subdistrik untuk dapat mengestimasi beban atas subdistrik untuk memiliki data yang mencukupi (29 dari 59 subdistirk sample memiliki kurang dari 20 respon). Maka, beban statistika 1990 dicocokkan pada jumlah total distrik 2004 dengan menggunakan perhitungan berikut:

- distrik_penyesuaian = (Total RT di tahun 2004 di distrik ini)/(Total 1990 beban di distrik ini)pps_beban_2004 = pps_beban_1990_penyesuan_kasar * distrik

145. Dengan memaksakan beban statistika untuk dicocokkan dengan jumlah rumah tangga distrik pada sensus 2004, beban statistika dinormalisasikan untuk dijumlahkan ke total rumah tangga pada tahun 2004 (194,943). Kesalahan-kesalahan terdapat dalam hasil diperhitungkan dengan menggunakan modul survey standar Stata.[9] Modul-modul ini menggunakan variabel desain survei (stratum, unit sampling primer dan beban sampling) untuk membuat estimasi beban dari total dan Taylor-series approximations dari kesalahan sampling. Estimasi kesalahan ini mengambil sampling acak dengan beban sampling yang tidak sama. Asumsi ini konservatif (yaitu, ia akan cenderung untuk menyepelekan kesalahan sampling) dengan mempertimbangkan beban statistika yang dihitung dengan metoda PPS yang dijelaskan sebelumnya.[10] File-file data elektronik yang digunakan untuk kalkulasi-kalkulasi ini tersedia di. http://www.hrdag.org/timor.

Penentuan tanggal RMS untuk analisa pemindahan

146. Survei ini menanyakan kepada responden kapan mereka pindah dari setiap lokasi selama periode 1974-1999. Ketika para responden tidak yakin akan tanggal pasti dari waktu pindah mereka, seringkali mereka mengidentifikasi tahun pindah mereka dengan titik tertentu dalam siklus pertanian atau apakah itu adalah musim kering atau hujan. Untuk setiap tanggal parsial atau musiman ini, kami menentukan pemindahan ke dalam kuartal di mana masa atau musim itu terjadi. Jika identifikasi tanggal parsial dapat terjadi di dalam lebih dari satu kuartal, ia ditetapkan secara acak ke dalam satu kuartal. Dari 2,024 proses pindah yang diartikan oleh responden sebagai kejadian pemindahan, 76.6% setidak-tidaknya diidentifikasikan ke dalam kuartal, dan 15.7% lainnya diidentifikasikan atas musim. Hanya 7.7% kejadian pemindahan diidentifikasikan atas tahun tanpa menyebutkan bulan.

Penyesuaian beban RMS untuk perkiraan angka kematian

147. Kalkulasi beban statistika mengasumsi bahwa kejadian-kejadian yang dilaporkan oleh setiap rumah tangga hanya bisa dilaporkan oleh rumah tangga tersebut. Asumsi ini ialah hasil dari beban yang merupakan resiprokal dari probabilitas sampling untuk setiap rumah tangga. Karena itu, apabila ada lebih dari satu rumah tangga yang dapat memberikan informasi tentang sebuah kematian tertentu, probabilitas sampling murni untuk kematian tersebut lebih besar dari probabilitas untuk satu rumah tangga. Kematian yang dilaporkan oleh responden survei melanggar asumsi laporan-tunggal-rumahtangga karena untuk setiap kematian, mungkin ada lebih dari satu rumah tangga yang dapat memberi laporan tentang kematian tersebut. Di antara 5,402 jumlah total kematian yang dilaporkan oleh responden, 545 dilaporkan lebih dari satu kali (laporan yang terduplikasi ini diidentifikasi dan dikeluarkan sebelum proses perkiraan). Duplikasi laporan yang terimplisit di dalam perkiraan beban survei diubah dengan menyesuaikan beban dengan cara yang dijelaskan di bawah ini.

148. Sebelum beban survey dapat digunakan untuk memperkirakan angka kematian total, they must be adjusted untuk menghitung jumlah rumah tangga yang merupakan responden potensial untuk setiap kematian. Yaitu, untuk setiap kematian, berapa anggota keluarga yang selamat sampai tahun 2003 untuk menjadi responden potensial dalam survey ini? Banyak

informasi yang dibutuhkan untuk perhitungan ini tersedia dalam survey ini karena anggota keluarga responden juga adalah anggota keluarga almarhum/almarhumah. Jumlah anggota keluarga yang selamat dari setiap korban *D* dikalkulasi berdasarkan anggota keluarga yang dilaporkan oleh responden *R* dengan menggunakan aturan-aturan sebagai berikut:

1. Jika *D* adalah orang tua dari *R*, jumlah anggota keluarga yang diharapkan selamat pada tahun 2003 adalah penjumlahan dari*:*

- Dianggap bahwa orang tua dari *D* adalah 25 tahun lebih tua dari *D* (atau 50 tahun lebih tua dari *R*, jika usia *D* tidak dilaporkan); gunakan probabilitas selamat kondisional spesifik-umur (yang dikalkulasi dari survey ini) untuk memperkirakan jumlah yang diharapkan jumlah orang tua yang diharapkan masih hidup pada tahun 2003
- Hitung saudara-saudara kandung *R* sebagai anak-anak dari *D*
- Dengan rata-rata terdekat tingkat kesuburan total adalah 5 sebelum tahun 1975, dianggap bahwa *D* mempunyai empat saudara dengan usia masing-masing (-4, -2, +2, +4) tahun dari usia *D* (jika usia *D* hilang, pasang usia *D terhadap usia R* + 25), calculate usia saudara kandung pada tahun 2003, dan kalikan masing-masing dengan kemuningkan bersyarat untuk selamat sampai usia tersebut, dan tambahkan keempat saudara kandung itu.

2. Jika *D* adalah seorang saudara dari *R*

- *Orang tua D adalah orang tua R*, langsung hitung orang yang selamat
- *Saudara R* adalah sudara *D*, langsung hitung orang yang selamat

149. Dianggap bahwa D mempunyai jumlah anak dewasa yang selamat yang sama dengan R.

3. Jika *D* adalah anak dari *R*

- *R* dan pasangannya adalah orang tua spouse are parents, langsung hitung orang yang selamat
- Anak-anak dewasa dari *R* adalah saudara-saudara kandung *D*, langsung hitung orang yang selamat
- Dianggap tidak ada anak dewasa yang selamat dari *D*.

150. Kalkulasi ini menghasilkan anggota keluarga dewasa yang diperkirakan selamat untuk setiap *D*, juga menunjukkan siapa dari anggota-anggota keluarga yang selamat ini tinggal dalam rumah tangga *R*, yang siapa yang tinggal di rumah tangga yang lain.

151. Untuk mengkonversi anggota keluarga dewasa yang diperkirakan selamat dari *D* ke dalam sebuah penyesuaian untuk beban sampling, jumlah anggota keluarga harus dikonversi ke dalam sebuah jumlah perkiraan rumah tangga di mana anggota keluarga tersebut tinggal. Terdapat kira-kira 0.5 anggota keluarga *D* (selain *R*) tinggal di rumah tangga *R*. Anggap bahwa rumah tangga lain di mana anggota keluarga *D* tinggal mempunyai jumlah pemusatan anggota keluarga per rumah tangga yang sama dengan rumah tangga *R* (yaitu 1.5 anggota per rumah tangga). Maka, apabila *D* mempunyai *L* anggota keluarga yang selamat yang tinggal di luar rumah tangga *R*, maka ada $a = 1 + L/1.5$ rumah tangga yang dapat memberikan informasi tentang *D*. Beban survey yang disesuaikan untuk kemungkinan laporan berganda atas *D* dengan membagi setiap beban sampling *D* dengan factor ini, *a*. Kalkulasi ini mengasumsikan bahwa rumah tangga responden potensial lainnya ada dalam kluster yang sama dengan *R*, atau bahwa mereka ada dalam kluster dengan kemungkinan sampling yang sama dalam kluster.

Analisis sensitifitas terhadap asumsi-asumsi dalam pengukuran kembali angka kematian (mortality re-weighting)

152. Ada sejumlah asumsi dalam pencocokan beban untuk perkiraan angka kematian, antara lain sebagai berikut:

- Perbedaan periode antara generasi-generasi (diasumsikan 25 tahun)
- Jumlah saudara kandung yang dimiliki orang tua responden (diasumsikan empat)
- Jarak kelahiran saudara kandung orang tua (diasumsikan dua tahun)
- Jumlah anak dewasa yang dimiliki saudara kandung responden (diasumsikan setara dengan anak-anak responden).

153. Asumsi-asumsi ini dites dengan menggunakan variasi-variasi berikut, dan total angka kematian tahunan dikalkulasi:

- Penjarakan antar-generasi bervariasi antara 18 dan 30 tahun
- Jumlah saudara dari orangtua responden diasumsikan telah bertambah menjadi enam
- Jarak kelahiran meningkat menjadi lima tahun antar saudara
- jumlah anak dewasa yang dimiliki saudara responden diasumsikan menjadi dua kali lipat dari anak-anak responden.

154. Untuk setiap varian perkiraan, total tahunan diuji (dengan menggunakan two-mean t-test) terhadap model utama. Tidak satu tahun pun di setiap model varian yang berbeda secara penting pada $p<0.05$. Nilai p minimum adalah 0.13, dan itu adalah sebuah outlier: nilai terendah kedua dari p adalah 0.23. Karena itu, perkiraan ini tidak sensitive secara mendasar terhadap asumsi mengenai struktur keluarga.

155. Meskipun perkiraan-perkiraan ini kaku terhadap asumsi-asumsi tentang struktur keluarga yang digunakan untuk mengestimasi jumlah anggota keluarga yang selamat yang dapat memberikan informasi tentang D, magnitude dari yang diperkirakan sensitif terhadap model yang digunakan untuk mentranformasi perkiraan anggota keluarga yang selamat untuk memperkirakan rumah tangga yang mengandung anggota keluarga. Perkiraan jumlah anggota keluarga yang selamat adalah L, dan perkiraan jumlah rumah tangga yang mempuyai anggota keluarga dari almarhum/almarhumah D, merupakan a, adalah $a = 1 + L/1.5$. Denominator 1.5 didapatkan dari rata-rata jumlah anggota keluarga untuk D (termasuk R) yang hidup di rumah tangga R (0.5).Memvariasikan rata-rata ini dari 0 sampai 3 (diasumsikan 1-4 anggota keluarga dewasa yang selamat per rumah tangga), meragamkan perkiraan hasil dari kematian total yang diperkirakan (oleh semua penyebab) dari -14.2% ke +19.6%. Dampak dari meragamkan model ini berkurang sepanjang waktu, dengan variasi terbesar ditemukan pada awal tahun 1972-1975 (-21%, +26%) dan variasi terkecil ditemukan pada tahun-tahun yang lebih kemudian yaitu 2001-2003 (-11%, +16.2%). Pengurangan ini konsisten sepanjang waktu.

156. Dengan adanya jumlah yang konstan dari anggota keluarga yang selamat, sedikit saja anggota keluarga yang selamat per rumah tangga sudah dapat berarti bahwa lebih banyak kemungkinan rumah tangga yang dilaporkan, bahwa ada kemungkinan perkiraan sampel yang tinggi per kematian yang dilaporkan, dan bahwa beban sampel per kematian yang dilaporkan yang rendah, dan karenanya perkiraan total kematian lebih rendah; lebih banyak orang dewasa per rumah tangga yang membalikkan logika ini.

157. Meskipun total perkiraan bervariasi dengan perubahan-perubahan dalam model-model yang mentransformasikan anggota-anggota keluarga ke dalam rumah tangga-rumah tangga,

polanya tetap sama. Koefisien korelasi untuk model utama ke model-model rendah (0) dan tinggi (3) di atas masing-masing adalah 0.99. Meskipun model anggota keluarga-per-rumah tangga memang mempengaruhi magnitude total dari kematian-kematian yang diperkirakan, model ini tidak mempengaruhi pola-pola perkiraam sepanjang waktu.

Multiple Systems Estimation (MSE): motivasi dan teori

158. Analisis survey adalah konservatif dalam pengertian bahwa analisa tersebut mengoreksi kemungkinan laporan berganda dengan mencocokan kematian di seluruh rumah tangga, dan karena ada penyesuaian terhadap beban sampling yang didasarkan pada perkiraan jumlah rumah tangga yang dapat memberikan informasi tentang setiap kematian. Walapun sejumlah kematian dapat saja dilaporkan oleh beberapa rumah tangga, ada juga kematian-kematian lain yang terjadi selama 1974-1999 di mana tidak ada anggota keluarga yang selamat di tahun 2003. Jika seluruh anggota keluarga meninggal selama periode mandat Komisi, mungkin saja tidak ada anggota keluarga kolineal yang dapat memberikan informasi di tahun 2003. Karena keterbatasan-keterbatasan ini, sebuah metode pengganti untuk memperkirakan total kematian dapat saja melakukan pengecekan pada survey yang diperkirakan.[*]

159. MSE menggunakan beberapa daftar populasi yang dikumpulkan secara terpisah yang tidak lengkap. Daftar-daftar ini mengidentifikasi kecocokan elemen-elemen yang sama sepanjang daftar-daftar itu untuk dapat memperkirakan jumlah elemen yang hilang dari seluruh daftar. Dalam proyek ini, kematian-kematian yang didokumentasikan dalam HRVD, RMS, dan GCD dicocokkan sepanjang ketiga system dengan menggunakan nama, tanggal kematian, lokasi kematian dan tanggal lahir.

160. Bentuk paling dasar dari tehnik ini adalah capture-tag-recapture, yang menggunakan hanya dua daftar.

161. Penjelasan tehnis tentang bagaimana sebuah perhitungan akan anggota yang tidak diketahui dari populasi dapat diperkiraan adalah sebagai berikut. Anggap saja ada kasus dari dua proyek P_1 (sebuah daftar individu-individu *A*) dan P_2 (sebuah daftar individu-individu *B*). Ada individu-individu *M* yang cocok di sepanjang daftar, dalam sebuah universe total individu-individu *N* (*N* tidak diketahui). Jika semua orang dalam universe *N* mempunyai kemungkinan yang sama untuk muncul dalam Daftar 1, maka kemungkinan seorang individu tertentu dilaporkan oleh P_1 adalah

162. *Pr(terdapat dalam daftar 1)* = $\frac{A}{}$

163. Demikian juga, jika semua orang dalam universe *N* mempunyai kemungkinan yang sama untuk muncul dalam Daftar 2, maka kemungkinan seorang individu tertentu dilaporkan oleh P_2 adalah

164. *Pr(terdapat dalam daftar 2)* = $\frac{B}{}$

165. Kemungkinan seorang individu tertentu terdapat/tertangkap dalam kedua daftar adalah

[*]Penjelasan ini sesuai 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. tersedia online pada http://shr.aaas.org/hrdag/peru

Pr(terdapat dalam daftar 1 dan daftar 2) = $\frac{M}{}$

166. Menurut definisi, kemungkinan dari sebuah peristiwa mengandung dua peristiwa yang terpisah adalah hasil dari kemungkinan independent itu. Kerena itu,

Pr(terdapat dalam daftar 1 dan 2) = Pr(terdapat dalam daftar 1) x Pr(terdapat dalam daftar 2)

167. Dimana $\frac{M}{}=\frac{A}{}\left(\frac{}{N}\right)$: dalam persamaan ini, jawaban untuk *N*. Rearranging the terms, $\frac{M}{}=\frac{B}{2}$ dan kemudian dikalikan dengan *N* , $M=\frac{}{N}$ dikalikan lagi dengan $M = B$, dan akhirnya dibagi dengan M menghasilkan $N=\frac{}{M}$. Catat bahwa dalam persamaan terakhir, jumlah kematian total *N* dapat diperkirakan dengan menggunakan total dari *A* dan *B* dan dari pencocokan antara mereka, *M*.

168. Ada banyak asumsi terimplisit dalam solusi ini. Sebagai contoh, tidak satu pun daftar yang mempunyai perorangan yang dilaporkan dua kali dan bahwa pencocokan antar daftar adalah akurat. Dalam proyek ini kedua asumsi ini dikendalikan selama pengolahan data sebagaimana dijelaskan dalam bagian pencocokan.

169. Asumsi lain yang inheren dalam model capture-tag-recapture lebih sulit ditangani. Pertama, metode ini mengasumsikan bahwa individu-individu tidak memasuki atau keluar dari universe ini selama proses penciptaan daftar-daftar itu, dan kedua bahwa daftar-daftar ini diseleksi secara acak dari populasi. Dalam proyek dokumentasi hak asasi manusia, asumsi pertama biasanya menjadi irrelevan karena dokumentasi terjadi secara retrospectif. Asumsi kedua tidak mungkin memuaskan, dan harus digantikan oleh asumsi bahwa perkiraan itu kaku terhadap proses-proses seleksi.

170. Asumsi yang lain adalah bahwa daftar-daftar itu independen, yaitu, bahwa kemungkinan bahwa seorang individual ada dalam daftar kedua adalah independen dari kemungkinan bahwa individual itu terdapat dalam daftar pertama. Asumsi terakhir adalah homogenitas: bahwa individu yang membentuk universe semuanya mempunyai kemungkinan yang sama untuk tertangkap dalam daftar.

171. Jika satu saja dari asumsi-asumsi ini tergganggu, metode capture-tag-recapture tidak akan menghasilkan sebuah estimasi yang memadai tentang besarnya populasi total. Jika ada lebih dari dua daftar dengan informasi yang memadai, masalah ketergantungan atau heterogenitas sering dapat diatasi melalui spesifikasi dan seleksi atas model-model yang tepat. Namun, dalam data untuk HRVD, RMS, dan GCD, hanya ada dua system yang dapat digunakan (RMS-GCD untuk kematian karena lapar atau kesakitan, dan HRVD-GCD untuk pembunuhan).[*] Pada dirinya sendiri perkiraan-perkiraan ini tidak akan memadai, tetapi dalam kombinasi dengan perkiraan RMS, perkiraan-perkiraan ini menyediakan informasi tambahan yang bermanfaat.

[*] Aplikasi awal dari estimasi system berganda terhadap estimasi demografis adalah oleh C Chandra Sekar dan 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. Sebuah diskusi menyeluruh tentang estimasi untuk pendekatan dual-system dan kesalahan kalkulasi yang mungkin berhubungan tersedia dalam Yvonne M M Bishop, Stephen E Fienberg dan Paul H. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. 1975. Untuk komentar tentang penggunaan metode-metode ini dalam analisis hak asasi manusia, lihat Fritz Scheuren, "History Corner," *The American Statistician*, Februari 2004.

Mengalokasikan GCD menurut jenis kematian

172. Data kuburan tidak memasukan cara kematian. Terdapat 89,894 kubur dengan setidak-tidaknya sebuah inisial awal (atau nama), sebuah nama belakang dan sebuah tahun antara 1972 dan 2003. Ada sejumlah 7,117 cocok baik dengan HRVD atau dengan RMS (atau keduanya), dan melalui kecocokan ini, cara kematian dapat diketahui dari catatan tentang cara yang cocok dengan kematian. Sisanya 82,717 catatan GCD perlu dialokasikan ke dalam empat kategori cara kematian (pembunuhan, kematian karena kelaparan atau kesakitan, kematian penempur, dan kematian lainnya). Dari RMS, proporsi kematian tahunan menurut keempat jenis ini ditunjukkan dalam Figure <number>, dibawah. Catat bahwa proporsi ini tidak mengikutsertakan kematian dimana cara kematian tidak diketahui (204 dari 3,235 kematian yang dilaporkan dalam RMS antara 1969 dan 2004 mempunyai cara kematian yang tidak diketahui).

Table 4 - Figure <number>: perkiraan proporsi kematian, menurut periode dan cara kematian

Periode	Pembunuhan	Kelaparan/Kesakitan	Penempur	Lain-lain
1972-1974	0.9%	95.9%	0.0%	3.2%
Marjin kesalahan	1.8%	5.1%	0.0%	4.9%
1975-1982	11.2%	83.0%	4.4%	1.4%
Marjin kesalahan	4.7%	5.1%	2.5%	0.6%
1983-1998	5.5%	86.5%	0.7%	7.2%
Marjin kesalahan	2.5%	3.7%	0.6%	2.5%
1999	16.2%	83.0%	0.4%	0.4%
Marjin kesalahan	10.2%	10.2%	0.8%	0.8%
2000-2003	3.5%	86.9%	0.8%	8.9%
Marjin kesalahan	3.1%	6.5%	1.6%	4.9%
Total	8.3%	85.1%	2.4%	4.3%
Marjin kesalahan	2.7%	3.1%	1.2%	1.2%

173. Proposi-proporsi ini digunakan untuk mengalokasikan catatan GCD yang belum tercocokkan ke dalam cara khusus kematian untuk digunakan di dalam kalkulasi MSE untuk setiap tahun: proporsi dari periode ini yang mengadung setiap tahun digunakan mengalokasikan kematian GCD dalam tahun itu. Marjin kesalahan dari alokasi dimasukkan dalam kesalahan yang diperkirakan untuk perkiraan MSE.

Analisis sensitifitas hilangnya pengetahuan social: penyesuaian-penyesuaian untuk perkiraan yang terlalu rendah

174. Survey ini menanyakan kepada responden tentang kematian orang tua, saudara kandung dan anak-anak mereka. Namun, beberapa kematian tidak menyisakan orang tua, saudara kandung atau anak-anak tetap hidup ketika survey ini dilaksanakan pada tahun 2004. Jika kematian-kematian telah terjadi lama di waktu lalu, bahkan anak-anak dari orang yang mati juga mati semuanya, tidak ada yang bisa melaporkan kematian-kematian tersebut. Dalam kasus lain keluarga yang kecil dapat saja menderita kematian semua anggota keluarga, sehingga tak seorangpun yang selamat untuk melaporkan kematian-kematian tersebut. Karena survey ini memperkirakan angka (atau tingkatan) kematian di masa lalu, perkiraan yang terlalu rendah yang merupakan akibat dari hilangnya pengetahuan sosial tentu saja menjadi sangat hebat. Namun,

bahkan dalam waktu yang belum terlalu lama (sebagai contoh, di tahun 2003 karena survey dilakukan tahun 2004), adalah tidak mungkin untuk mendokumentasikan sejumlah kematian yang tidak mempunyai seorang pun yang selamat. Sebagai contoh, orang-orang yang tidak mempunyai orang tua, saudara kandung atau anak-anak, yang meninggal di tahun 2003 tidak dapat dilaporkan dalam survey ini.

175. Tingkat kematian kasar (per 1,000 orang) adalah sebuah perkiraan tentang berapa banyak orang meninggal, secara total, per tahun. Tingkat kematian kasar ini adalah standard demografik dan indicator kesehatan, yang biasanya diperkiraan dengan metode-metode tidak langsung dengan menggunakan catatan-catatan sensus. Untuk Timor-Leste, tingkat ini sulit diperkirakan karena kualitas data sensus tahun 1980 dan 1990 masih dipertentangkan.[11] CDR yang diperkirakan oleh Biro Amerika Serikat untuk Sensus bagi Timor-Leste ditunjukkan untuk tahun 1990-2004. Tingkat keseluruhan untuk Indonesia ditunjukkan dalam tahun 1983. Perkiraan yang ditunjukkan untuk tahun 1971 yang berasal dari pemerintah Indonesia mengklaim bahwa di seluruh Indonesia antara tahun 1971 dan 1990, CDR menurun sampai 45%; perkiraan tahun 1971 yang ditunjukkan di sini adalah perkiraan tahun 1990 untuk Timor-Leste digelembungkan oleh factor ini. Suatu CDR yang sudah diperhitungkan juga ditunjukkan dengan secara linear disisipkan antara perkiraan tahun 1971 dan perkiraan tahun 1990-2004.

176. Selain perkiraan CDR, CDR dari RMS Komisi juga ditunjukkan. Perkiraan ini adalah total perkiraan kematian dibagi dengan perkiraan populasi untuk tahun itu (dikali 1,000). Ada beberapa pengamatan yang dibuat mengenai grafik ini. Pertama, CDR yang diperkirakan oleh Biro Amerika Serikat adalah dalam interval yang dipercaya dari CDR yang diperkirakan oleh RMS yang dimulai dari tahun 1993. Pada tahun 2003 interval yang dipercaya dari RMS CDR (4.2 – 6.6) mengadung perkiraan Biro Amerika Serikat (6.4), sebagaimana ditunjukkan dalam grafik by the capped spike at the end of the CAVR line. That is, meskipun RMS sangat rendah memperkirakan tingkat kematian dalam tahun-tahun damai yang “normal” 1972-1974, sejak pertengahan tahun 1990-an, RMS sesuai dengan hasil yang didapatkan melalui metode-metode tidak langsung yang digunakan oleh Biro Amerika Serikat. Pengamatan ini sesuai dengan gagasan bahwa perkiraan RMS mempunyai bias penurunan yang meningkat ke masa lalu.

177. Selama tahun-tahun di mana catatan-catatan historis menyatakan bahwa kematian yang sangat berlebihan terjadi, interpolasi linear CDR memperkiraan terlalu rendah kematian-kematian tersebut. Tahun-tahun ini termasuk tahun 1975-1979 dan 1999. Hal ini sesuai dengan arti literal dari kematian “berlebihan”. (Tidak ada CDR berbasis sensus yang diperkirakan untuk periode tahun 1975-1979). Melihat lebih jauh ke belakang, CDR berbasis survey menangkap sedikit penurunan dari total CDR (sebuah grafik yang sama dapat digambarkan untuk MSE yang diperkirakan atas waktu, dengan hasil yang sama).

178. Untuk menyesuaikan RMS, kematian-kematian yang tidak diketahui karena hilangnya pengetahuan sosial harus di perkiraan sepanjang waktu. Model yang digunakan adalah sebagai berikut:

- Angka kematian diperkirakan oleh CDR dan population yang sudah diperhitungkan untuk setiap tahun diperkiraan (kematian-kematian CDR), ditunjukkan sebagi sebuah rate dalam Figure {g_cdrs.pdf};
- pecahan dari kematian-kematian CDR yang terjadi karena kelaparan dan kesakitan diperkirakan dengan menggunakan pecahan dari semua kematian yang dilaporkan dalam survey yang dikarenakan oleh kelaparan dan kesakitan (serupa dengan alokasi yang digunakan untuk data GCD yang tidak tercocokan). Dalam survey ini, pecahan mean (dan median) dari semua kematian (selama bertahun-tahun) yang dihubungkan dengan kelaparan dan kesakitan adalah 0.80, dan 50% dari semua tahun adalah antara jenjang 0.754 – 0.846;
- rasio kematian-kematian terestimasi terhadap kematian_CDR dikalkulasikan untuk tahun-tahun damai (1972-1974 dan 2002-2003); ini adalah pecahan dari "kematian-kematian yang dapat diingat" yang disebut "pecahan memory;"
- Pecahan memory untuk 1975-2001 diperkirakan oleh interpolasi linear dengan menggunakan persamaan berikut
- Pecahan memory terperkirakan (MSE) = -39.1 + 0.0200*tahun
- Pecahan memory terperkirakan (RMS) = -43.9 + 0.0224*tahun

- Pecahan memory untuk MSE berkisar dari 0.241-0.936, di mana untuk RMS, pecahan memory tersebut berkisar dari 0.228 to 0.846. Perbedaan ini mempunyai pengaruh yang luar biasa pada hasil/keluaran.
- Perkiraan tersesuaikan dikalkulasi sebagai perkiraan asli dibagi dengan pecahan memory untuk setiap tahun.

179. Perkiraan tersesuaikan yang disampaikan di bawah ini dalam Figures {g_huil_xs_mse.pdf} dan {g_huil_xs_rms.pdf}. Catat bahwa dalam kedua grafik itu, perkiraan kasar dan perkiraan yang tersesuaikan bertemu saat mendekati tahun 2003. Dampak dari pecahan memory yang tinggi untuk MSE yang berhubungan dengan RMS jelas dalam perkiraan total kematian lebih dari baseline CDR: perkiraan tersesuaikan MSE adalah 104,000 kematian sedangkan perkiraan tersesuaikan RMS adalah 183,300 kematian.

180. Kedua perkiraan ini bergantung kepada sejumlah asumsi, termasuk asumsi tentang bentuk penurunan CDR sejak awal 1970-an sampai akhir 1990-an dan tentang sifat dari hilangnya memory sosial. Perubahan secara harus namun tidak linear dalam hilangnya memory sosial (entah melengkung naik atau melengkung turun) tidak akan merubah perkiraan secara berarti. Namun, jika perkiraan terlalu rendah dalam MSE dan RMS yang dikarenakan oleh kehilangan memory sosial adalah tidak berkelanjutan atau sebaliknya sangat berbeda untuk tahun 1972-1974 yang berhubungan dengan tahun-tahun puncak 1975-1979, penyesuaian yang dilakukan disini tidak akan mengoreksi secara memadai terhadap perkiraan yang terlalu rendah itu. Kedua model ini bergantung pada kalkulasi CDR dari data sensus tahun 1980 dan 1990 dan metode-metode tidak langsung yang digunakan oleh Biro Sensus Amerika Serikat. Ada kesalahan sampling dan non-sampling yang tidak tergambarkan dalam grafik-grafik atau statisitik, tetapi kesalahan-kesalahan itu tentu saja mendasar.

181. Namun, model-model ini mempunyai keuntungan menunjukkan bahwa dalam penyesuaian, perkiraan total kematian tahunan yang dikarenakan oleh kelaparan atau kesakitan lebih dekat kecocokannya dengan baseline kematian-kematian CDR untuk periode sebelum invasi (1972-1974) dan untuk periode 1984-1998.

182. Ada sejumlah alasan untuk memilih perkiraan MSE daripada perkiraan RMS. Meskipun RMS lebih dekat kecocokannya dengan perkiraan kematian-kematian CDR dalam tahun-tahun setelah pendudukan yang mendekati waktu damai, 2002-2003, MSE lebih dekat kecocokannya

dengan perkiraan kematian total CDR sebelum pendudukan. Untuk tujuan perkiraan ini, periode yang paling relevan adalah periode 1975-1979, dan pilihan perkiraan harus dituntun oleh kesesuaian yang terbaik segera sebelum periode ini. Alasan kedua untuk memilih MSE adalah bahwa MSE didasarkan pada data yang sangat lebih banyak dari RMS sendiri: MSE menggunakan data GCD disamping data RMS.

183. Kesimpulan yang paling kuat yang dapat dibuat adalah bahwa perkiraan RMS dan MSE yang tak tersesuaikan terlalu rendah. Bagian 6: Profil Pelanggaran Hak Asasi Manusia, menyediakan sebuah pemeriksaan terhadap dukungan statistic untuk temuan-temuan dalam hubungan dengan jumlah pelanggaran fatal selama periode mandat Komisi.

---

[1] Regulasi UNTAET 2001/10 Seksi 13.1(a)(i).

[2] Regulasi UNTAET 2001/10 Seksi 13.1(a)(i).

[3] Regulasi UNTAET 2001/10 Seksi 13.1(a)(i).

[4] Regulasi UNTAET 2001/10 Seksi 13.1(a)(ii).

[5] Regulasi UNTAET 2001/10 Seksi 13.1(a)(iv).

[6] Regulasi UNTAET 2001/10 Seksi 13.1(d).

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

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

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

[10] Donna Brogan, "Sampling error estimation for survey data," dalam 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.

[11] lihat misalnya, 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), h. 585-597.