H i C N **H**ouseholds **i**n **C**onflict **N**etwork

The Institute of Development Studies - at the University of Sussex - Falmer - Brighton - BN1 9RE

www.hicn.org

The Bosnian Book of Dead: Assessment of the Database (Full Report)

Patrick Ball, Ewa Tabeau and Philip Verwimp

HiCN Research Design Note 5

17 June 2007

THE BOSNIAN BOOK OF DEAD: ASSESSMENT OF THE DATABASE (FULL REPORT)[1]

Patrick Ball, Ewa Tabeau and Philip Verwimp[2]

17 June 2007

Table of Contents:

Acknowledgement

This report summarizes results from a research project we conducted at the invitation of Mirsad Tokača, the president of the Research and Documentatiojn Centre in Sarajevo and two Embassies, of Norway and Switzerland, in Sarajevo. We want to thank Mr. Tokača and the staff of the RDC for their efficiency and dedication in supporting us with all data, documentation and clarifications needed for our work. We want to thank the two Embassies for inviting us for studying this important and unique database. We are also very grateful to our employers, the Benetech Initiative (USA), the Office of the Prosecutor of the United Nations International Criminal Tribunal for the former Yugoslavia (Netherlands), and the Microcon (a European Union research group), for making it possible for us to complete this work.

[1] The views expressed in this report are of the authors alone and do not necessarily correspond to the opinions of the organizations the authors are affiliated with. In particular, the opinions expressed in this report do not necessarily represent those of the Office of the Prosecutor of the International Criminal Tribunal for the former Yugoslavia, or the United Nations.
[2] Authors'order is alphabetic.

EXECUTIVE SUMMARY

This Executive Summary relates to the expert report "Bosnian Book of Dead: Assessment of the Database", by Patrick Ball, Ewa Tabeau and Philip Verwimp. The report concerns the database "Bosnian Book of Dead" (hereafter: the BBD Database), known as well as the Population Loss Project of the Research and Documentation Centre in Sarajevo (RDC) presided by Mirsad Tokača. The BBD Database is a Bosnia and Herzegovina (BH)-wide database on 1992-95 war-related deaths. As of July 2006, it contained 96,985 cases representing individual victims of war. We obtained a copy of the database from Mirsad Tokača, the president of RDC, at a meeting in Sarajevo in July 2006, at which the assessment project was initiated. We conducted this project at the invitation of Mirsad Tokača himself and two embassies, of Norway and Switzerland, which were seeking a validation of this work.[3]

The report is composed of this Executive Summary and a Technical Report. Experts' profiles are available from the Annex to this report. Executive Summary contains general findings and recommendations for improvements and uses of the database. Details of our work are available from the Technical Report.

In our work we investigated three areas of the BBD Database:

1. Data problems such as:
    - Errors, data cleaning, outliers, and missing values
    - Consistency of reporting on victim details (e.g. names and other personal details, ethnicity or civilian-military status) and death characteristics (i.e. time, place and cause of death)
    - Complete versus less complete records and monitoring the loss of information related to deficiencies in the completeness.
2. Preservation of the original information in the database
    - Assessment of the "active" records created by combining pieces of information from overlapping sources in relation to these original sources;
    - Sources used for the BBD database
3. The coverage of the BBD database:
    - Overview of the BBD procedures for checking for and elimination of duplicates;

Here are our major findings:

General findings

[3] We completed our assessment on request of those who invited us to this project, and also for a broader audience; namely for all others interested in victims' aspects of the 1992-95 war in Bosnia and Herzegovina. This could be the families of the deceased who might take interest in the BBD Database, historians tracing the truth about the Bosnian war, politicians that have victims' issues on their agendas, NGOs that work on prevention of human rights violations in the Balkans region, observers of the reconciliation process in the region, international and national courts prosecuting individuals responsible for violations of the International Humanitarian Law and the Law of War in the 1992-95 conflicts in Bosnia and Herzegovina, and the media. All those that intend consulting this largest existing database on victims of the 1992-95 war in Bosnia for their personal interest, work or research, will find this assessment useful and instructive.

- The BBD contains 96,895 cases (or records), each related to one victim that was killed, died another way in war-related circumstances, or disappeared during the war. The 96,895 cases are called "active" and represent those records of the overall total of all collected cases (246,736), which have been approved as final. Many active records were reported in several sources, pieces of which are now contained in these records.
- Many consider the number of 96,895 as the overall total of victims of the 1992-95 war in Bosnia, which is not correct. For several reasons, this number should be seen as an approximation of a minimum and not as the complete total.
- The BBD statistics on victims are obtained from information collected mainly from individual informants, such as eye witnesses, close relatives, friends, neighbours etc. that provided this information voluntarily, or from overall sources on war-related victims, such as press reports, books, missing person's lists, NGOs, government sources etc. No standardized documents were required to prove statements of the respondents. For these reasons there might be some inconsistent and less reliable records included in the BBD as well.
- Even though it is the largest existing database on Bosnian war victims, the BBD should not be used alone but together with other sources on war victims or on incidents and episodes of the war; this will prevent from producing biased statistics and historically incorrect pictures, and help avoiding misinformation of the audience.
- The fact that including new cases brings only marginal improvements indicates that the most cases have already been placed in the database.
- All computerized records, those marked as "active" and those "inactive", have been assessed in our project. The active records were examined in relation to the data quality, incompleteness of reporting, and preservation of the original documented information about victims and their suffering in the computerized material. The entire database (active and inactive records) was used in analysis of duplicates.

Data quality and incompleteness of reporting

- Two groups of items were inspected on the first place: personal identification items and event (i.e. death or disappearance) identification items.
- Only a few problems were encountered within these two groups.
- Both groups were only slightly affected by data entry errors, or misplaced information, and more impact was seen for missing values.
- Except for missing values, all other deficiencies (excluding duplicates which are discussed separately) can be seen as extremely minor; many of them can be easily corrected by studying the records in the database and/or checking in the original source material what actually is wrong.
- Missing values are not a database problem. Missing values are a reporting problem; these were the informants that were unable or not requested to provide certain pieces of information to the BBD developers which resulted in incompleteness of certain data items in the database.
- Because of the missing values, not all of the active records (96,895) could be equally treated. About 85% of active records were relatively complete (82,257) and about 15% of active records were concluded less complete (14,638). The less complete records, which we identified as records of mainly civilians, women, and of "Other" ethnicity, can and should be improved and their value revised accordingly.

- The most frequent deficiency of BBD records is the missing year (and date) of birth (9,430 or 9.7%), the second most frequent is the missing year (and date) of death (7,428 or 7.7%), and the third most prominent deficiency are the duplicates (at least 1,060 or 1.1%; likely more).[4] Almost all decisions on declaring records less complete were related to these three shortcomings. In relative terms, the scale of these weaknesses is small, however.
- About 77.5% of less complete records (11,342 out of 14,638) are characterised by having a single deficiency (such as one missing data item). About 22.5% of less complete records are deficient on two or more items (3,296). This confirms the observation that deficient records tend to have missing values on one dimension only which is easy to repair.
- Records marked as complete can be relatively safely used in analysis. Statistics obtained based on complete records are the "minimum" or "at least" numbers and can be documented by relatively complete data from the database.
- Ethnicity is available for practically all records in the database (0.4% missing), thus the availability is not an issue. Moreover, the quality of reporting is clearly uniform among the ethnic groups and practically no ethnicity-related bias is present.
- Regarding the civilian-military status, called in the database "Status in War", it is reported on the basis of official military lists and other relevant sources. The RDC staff has visited 366 military and civilian cemeteries, gathered books and photos from military memorials, and collected a number of other appropriate sources that were all used to decide about the status of individuals. "Status in War" was obviously strictly based on the available sources and no arbitrary decisions were made with respect to the coding.
- Thus, "Status in War" merely shows the reporting of victims in military versus civilian sources. As of now, about 40% of victims are reported as civilians and 60% as soldiers (including the policemen). Among the complete records, the respective fractions are 36 and 64 percent.
- During our BBD project and other studies related to the registration of victims of war it came to our attention that some victims reported as soldiers according to official military lists, would be as well claimed civilians in civilian sources, and vice-versa. In particular, some military records could have been created by authorities in response to applications from the relatives of the deceased for the post-mortal benefits after the deceased. Secondly, some families might have found it honourable to bury their deceased among the defenders in military cemeteries or to publish their names on defenders' lists, even if the actual circumstances of death were not necessarily directly related to combat. These practices likely lead to over-reporting of soldiers and under-reporting of civilians in the sources. In consequence of these and other similar practices, civilians are in our opinion underrepresented in "Status in War".

[4] 1,060 is a minimum number of duplicates in the database. More duplications were found and are discussed in our Technical Report. Here the minimum is used in the context of the final completeness criterion. The declarations of cases complete or less complete will have to be updated by the authors of BBD after they will clean the duplicates.

- The civilian victims are also underrepresented among the complete records in the BBD as the drop-out ratio[5] for civilians (24.1%) is much higher than for soldiers (9.1%) or policemen (4.4%). Relatively many more civilians have been marked as less complete records whereas records of militaries were generally more complete.
- Improving the records of civilians is now a high priority and the analysis of the BBD data by "Status in War" should be postponed until a better data on civilians will become available. Especially, the military records reported as well in civilian sources should be reviewed and possibly revised. Secondly, the completeness of those records of civilians that are now marked as less complete should be enhanced.
- It is important to emphasize that "Status in War" does not provide correct insights in relation to victims of combat versus non-combat situations. Neither does it inform about legitimate victims of violations of the International Humanitarian Law or the Law of War. "Status in War" is a simple measure of whether or not a person was a member of a military/police formation at the time of death, (or generally was a defender), or a civilian. As such it offers a good basis for a further more specific investigation into issues related to involvement of victims in combat or proportionality of civilian losses.
- Being clearly aware of the above mentioned differences, the authors of the BBD have made an attempt to shed some light on combatants versus non-combatants issues. At this stage, however, this attempt cannot be concluded successful as the fraction of missing values on the combatant status ("Suffering-Level 2") is about 96% which disqualifies using this item in any analysis at all.
- The same is true in relation to another item, "Cause of Death – Military Formation", intended to describe which military formation caused deaths of victims. Again the 85% of values on this item are missing and this prevents from using this item at this stage.
- Also the mass grave part of the BBD Database (about 2,217 victims reported as exhumed from mass graves), is largely incomplete. The overall total of human remains exhumed from the mass graves (5 or more sets of remains) in Bosnia and Herzegovina was reported as 10,790 at the end of 2005 by the Federal Commission for Tracing Missing Persons. The overall total of identified persons was lower, (due to commingled remains and not-yet-available identifications), and equalled 8,724 persons. These numbers do not include statistics from the RS authorities and not from the Croat component of the Federal Commission. At present, the statistics are much higher. Noteworthy, the BBD team stressed to us that at this stage the exhumations part is not meant for analysis and has a supportive role in the database.

Preservation of the original information about victims

- In order to investigate the preservation, a special project was run in the RDC archive. In this project a random sample of 50 active cases, each case representing one individual, were selected from the BBD Database (0.05% of active cases) and 21 major data items (i.e. variables) available for every case were compared with the underlying documentation stored in the BBD archive. Leading questions were the

[5] The drop-out ratio is a simple measure of the relationship between the number of less complete records and the overall number of all records in a given category. The ratio shows the percentage of less complete records in the respective total.

following: what sources were used? What was the coding protocol? What decisions were made in the event of conflicting evidence between sources? How many errors and inconsistencies were there as compared with the original sources?

- Based on the above sample we concluded that out of the total of 21 investigated variables for each case, on average 17.3 variables were available per case, with the 95% confidence interval from 16.7 to 17.8. Note again, missing variables is not a database issue; it's an informant-related problem that can only be solved by comparing overlapping sources and improving the records accordingly which the BBD team has been doing all the time.
- Out of the available 17.3 variables, 17.1 were correctly computerized and 0.2 were errors. The error level is negligible.
- The mistakes include a duplicate, which is most serious. The other mistakes were minor, such as an inconsistent code for the municipality of the victim or a different code for the cause of death. Importantly, the team in the Centre is very aware of the remaining inaccuracies (such as the entry of the dates) and is constantly improving and updating the data base.

Duplicates

- Undetected duplicates, and the management of duplication within the database, may pose a more serious problem, especially with respect to the possibility of estimating the overall (unknown) total number of deaths during the war.
- Records marked as valid[6] can be relatively safely used in analysis. Statistics obtained based on valid records are the minimum or "at least" numbers and can be documented by relatively complete data from the database. However, we caution that comparisons among categories may be biased due to uneven rates at which deaths were reported; more detail is provided in Chapter 3 of Technical Report for different periods, different regions, combatant status, or different ethnicities. Such comparisons must be made using estimated totals which correct for under-reporting.

Any source on victims of an armed conflict is incomplete and deficient in many ways, which is a result of chaotic and traumatic circumstances of these deaths, the presence of conflict, and the fact that the functioning of the statistical institutions officially responsible for taking death records in the population and others collecting data on deaths is usually far from being normal in conflict situations. The BBD is not an exception among sources on war victims and must be expected to be incomplete and deficient too.

Yet the overall conclusion of our project is that the level of incompleteness and deficiency in the BBD Database is low and fairly acceptable. The database is a remarkable achievement of all those who have worked on this project. Further activities related to this source should concentrate on improving the quality of information and on enlarging its size by checking the records not yet marked as active and complete, which task although time-consuming is certainly feasible. A validation of the BBD Database with other sources such as for example, the 1991 Population Census, reporting on the Bosnian population at the outbreak of the 1992-

[6] In this context, "valid" are those records that are not duplicated.

95 war, or the FBH and RS Mortality Databases 1992-95 (referred to below), would be most desired.

In this context it is useful to note that sources on victims of the Bosnian war are generally extensive and include, for example:

- The FBH 1992-95 Mortality Database established in 2002 by the Federal Institute for Statistics in Sarajevo. (About 25,000 war-related and 50,000 natural death records).
- The RS 1992-95 Mortality Database compiled in 2005 by the Statistical Office of Republika Srpska in Banja Luka. (About 16,000 war-related and 50,000 natural death records).
- The ICRC list of missing persons. (About 22,000 records).
- Several other lists of missing persons including those by the FBH and RS Commissions for Tracing Missing Persons, another one by the International Commission for Missing Persons (ICMP) in Sarajevo, and several lists published locally (like for Prijedor and other municipalities).
- Official military lists of fallen soldiers and military and police personnel of the FBH and RS Ministries of Defence. (About 50,000 records)
- Records of the exhumed and identified persons in possession of the FBH and RS Commissions for Tracing Missing Persons, and of the ICMP. (The persons identified through the DNA matching methodology alone amounted recently to at least 8,000 individuals in Bosnia).
- Sarajevo Household Survey of mid-1994. (About 6,000 war-related deaths in Sarajevo until mid-1994).
- Many other lists by various NGOs.
- And of course, there is the Bosnian Book of Dead Database.

Each of the above sources is indispensable in answering specific questions meant to be answered by this given source. However, when it comes to statistics on victimization of a war, none of the above sources, if used alone, can be seen as sufficient. None of them can be then considered complete and unbiased with respect to statistics on victims of the 1992-95 war in Bosnia. The BBD is by far the largest and most complete source in this context. But the best approximations of the truth will be always obtained from results coming from many different sources and many different methodological approaches.

Having studied the 2006 version of the BBD extensively for the needs of this assessment and realizing a striking improvement of the 2006 version when compared with earlier versions of the database, we are happy to be able to recommend the use of the BBD for the following purposes:

- Advancing the reconciliation process in Bosnia and Herzegovina by displaying transparent and methodologically correct statistics on victims of BiH war.
- Propagating the approach and methodology used for the establishment of BBD. When presenting statistics, stressing the need of distinguishing between the <u>minimum numbers</u> and more complete <u>estimates.</u>

- Propagating comparisons of BBD with other sources on victims and additional sources on incidents and episodes of the war for the purpose of a better insight into the historical truth.
- Using the BBD Database for education of young researchers who can apply this knowledge in their careers.
- Using the database for lead purposes in investigative stages of trial preparation in international and/or national courts for IHL violations.
- Using the BBD database for academic research purposes, including expert analysis and testimonies for judicial proceedings.

The database is a unique and valuable source and deserves a prominent place among sources on victimization of the 1992-95 war in Bosnia and Herzegovina.

TECHNICAL REPORT

PART I: COMPLETENESS OF THE DATABASE

1.1 INTRODUCTION

The database "Bosnian Book of Dead" (BBD), known as well as the Population Loss Project, is a Bosnia and Herzegovina (BH)-wide 1992-95 war-related deaths database. The consultancy team obtained it from Mirsad Tokača, the president of the Research and Documentation Centre in Sarajevo (RDC; or the Centre), in July 2006, during an expert meeting there on 30th June to 1st July 2006. The Centre is the successor of the BH State Commission for Gathering Facts on War Crimes. The Commission operated in 1992-1995 and ceased its existence approximately two years after the war ended.

The intended coverage of the BBD database is the entire country and the entire conflict period 1992-95. The current version of the database is almost complete, meaning that marginal numbers of cases can be probably still found and added resulting in a diminishing improvement. It is the largest existing database on war-related deaths of both civilians and soldiers for Bosnia. As of July 2006, the total number of records is 246,736. However, only a fraction of this total is marked as active records, i.e. 96,895.[7] The remaining records were not flagged as active due to incompleteness, other deficiencies or duplicates.

Items included on the original CD ROM handed over to the consultants in July 2006 are the following:

a. "Victims": the main data table (in "txt" format) with the BBD records (246,736 entries, of which 96,895 marked as active). A related record description was provided as well.
b. Eight code books: eight files (in "txt" format) containing codes and their meanings for the following data items: "Mass Graves", "Military Formation at Death", "Military Formation", "Municipality", "Nation", "Religion", "Status in War", "Suffering".
c. PDF document (348 pages): "List of Sources" containing listing of sources used for the database.

An assessment of the quality of this original material (96,895 active records only) is the subject of this report. The final outcome of the assessment project was meant to be provided back to the authors of the database and to their sponsors, the Embassies of Norway and Switzerland, as a feedback that they could use in the future for further decisions related to the BBD project.

[7] The active records are those checked, corrected and approved by the authors of BBD. The remaining records consist mainly of those "checked and rejected"; some records might be still to check. Thus, the overall total of active records reported in the July 2006 version of BBD, i.e. 96,895, should be seen as a minimum number that will further slightly increase (not much though). Two more records might be active as well, but I excluded them due to inconsistent flagging.

From the start, we note that any source on victims of an armed conflict is incomplete and deficient in many ways, which is a result of chaotic and traumatic circumstances of these deaths, the presence of conflict, and the fact that the functioning of the statistical institutions officially responsible for taking death records in the population is usually far from being normal in conflict situations. Thus, the BBD Database as well – being one of the sources on conflict victims – must be expected to be incomplete and deficient. The role of the consultants was concentrated on concluding the level of incompleteness, major deficiencies of the data, and advising on the usefulness of this source.

Part I of this report summarises main data problems encountered in the course of the assessment project and ways of dealing with these problems. It also contains a few research results from the active records of the BBD Database. The part consists of 7 sections:

1.2 BOSNIAN BOOK OF DEAD: BACKGROUND INFORMATION

The Bosnian Book of Dead is the outcome of the project "Population Losses, 1992-95", conducted by the Research and Documentation Centre (RDC) in Sarajevo. The objective of this project is establishing a country-wide database covering the victims of the Bosnian war. Sources used for the BBD include witness statements[8], existing electronic lists, lists from books, reports, and press articles, names from grave tombs, newspaper memorials, other newspapers records (single or lists), government sources, microfilms etc. More than 7,000 witnesses testified so far and in total thousands of different sources were used (personal communication of the consultants with Mirsad Tokača, Sarajevo, 1 July 2006). All these sources are summarized in the document "List of Sources" which we studied as part of our assessment.

According to Mr. Tokača (*ibid*), the BBD project re-started at full speed in October 2003 by taking the MAG[9] mortality database and other computerized lists of victims as a starting point. In April 2004 the BBD contained 39,527 active records and 86,369 of such records in August 2004. In July 2006, the overall number of active records was 96,895, which represents a great increase within a short period of time. These were numbers of checked unique records. The overall number of entries in the database was much higher and equalled 246,736 as of July

[8] Eye witness statements were collected not necessarily according to investigative procedures. The RDC does not pretend they used the same methods as legal institutions do, but records were accepted only from eye witnesses, relatives, neighbours, and close friends.
[9] MAG stands for Muslims against Genocide, a non-governmental organization from Sarajevo. They do not exist anymore at present.

2006. A majority of these records were not marked as active due to various shortcomings (mainly duplicates). The project has six regional components distinguished according to the main conflict episodes during the Bosnian war:

- Eastern Bosnia
- Bosanska Krajina
- North Eastern Bosnia
- Sarajevo – Central Bosnia
- Herzegovina
- The remainder of Bosnia

On several occasions, the RDC produced preliminary statistics according to the above-mentioned regions and municipalities in the period between 1992 and 1995. As we will see, the reliability of these preliminary results based on all active records, although generally high, can be sometimes less than 100 percent due to various shortcomings of the data discussed in this report. The reliability can be improved by presenting results as minimum numbers rather than complete statistics. There are some problems with comparing minimum numbers of deaths across social categories, as will be seen in Part III.

Even though the RDC do not use the same methods of data collection and verification as legal institutions do, the material they accumulated is very useful as a historical record of demographic losses during the war. As such, the BBD Database has the potential of an extensive source material to be used by the legal institutions, such as international and/or national criminal courts, in prosecution of those responsible for violations of the International Criminal Law. In particular, this database could be used in investigative stages of trial preparation when lists of victims obtained from the database would provide leads for a further investigation. Last but not least, the database is an important step towards the closure of the war time losses and in advancing the reconciliation process in Bosnia and Herzegovina.

1.3 THE DATABASE

The original data tables (i.e. "Victims" table, code books and record description) were used in establishing an Access database called "BBD 2006.mdb" (hereafter: BBD Database). The original text files were converted to the Access format and properly organized. The Access database contains two main data tables:

- "Victims": original data converted to the MS Access format; 246,736 records,
- "Active Records": 96,895 records marked as "active" in "Victims"; used in the assessment summarised in this report.

In addition to these tables, the BBD Database also includes the eight original code books with an English translation of the codes.

At this stage it is important to stress that the list of sources, (several thousands of all kinds of informants, reports and publications reported in a separate PDF document called "List of

Sources"; 384 pages), is not available from the records in the Victims table. The Victims table contains only a very general link to groups of sources. In the actual system operating from the RDC in Sarajevo, the individual sources can be most certainly linked with cases in the database but the 2006 version of the BBD data the consultants have does not allow this. This is a serious deficiency from the investigative point of view. Chapter X provides additional elements on the treatment of the source material by the consultants as well as by RDC.

The list of data items available from the original Victims table is shown below in Table 1.

Table 1. List of Items Included in the BBD Database.

Data Item	Type	Description
Code of victim	Number	Record ID
Active	Text	Marker of active records
Last, First Name	Text	Names (first and family)
Fathers name	Text	Father's name
JMBG	Text	Maticni broj
Date of birth	Date	Date of birth
Place of birth	Text	Place of birth
Municipality (birth)	Number	Municipality of birth
Address	Text	Address (of residence?)
Code of nation	Number	Ethnicity (coded)
Code of religion	Number	Religion (coded)
Sex	Text	Sex
Occupation	Text	Occupation
Marital status	Text	Marital status
Criminal record	Text	Criminal record
Military formation	Number	Military formation
Date of suffering	Date	Date of suffering (here: death)
Code of suffering	Number	Type of suffering (here: cause of death)
Year of suffering	Text	Year of suffering (here: year of death)
Mass grave	Number	Mass grave (where the remains were exhumed)
Education	Text	Education of the victim
Status in war	Number	Status in war (here: intended as civilian-military status at death)
Grave location	Text	Grave location (here: where the person is buried)
Municipality of residence	Number	Municipality of residence (coded)
Municipality of suffering	Number	Municipality of suffering (coded)
Citizenship	Text	Citizenship
Dead	Text	Dead (here: whether confirmed death?)
Code of data origin	Number	Code of data origin (here: sources and/or informants; coded)
Code of suffering (level 2)	Number	Code of suffering (level 2; here: intended as legible or illegible victim of war)
Cause of death (Military Formation)	Number	Cause of death (Military formation) - here Military Formation at death

The assessment project focused on the following activities:

- reviewing all original items and inspection of errors,
- cleaning obvious deficiencies that could be cleaned without using additional sources or analysis of data (e.g. spelling errors or misplaced information),
- cleaning of deficiencies involving comparisons of items (i.e. checking logical links between items; e.g. comparisons of dates, such as DoB and DoD, or DoB with JMBG etc.),

- re-coding items using code books,
- duplicate checks and marking duplicated records for removal,
- checking availability of items needed to uniquely identify victims
- checking availability of items needed to uniquely identify the death or disappearance of victims and circumstances of these events,
- checking availability of items needed to describe victim's status as a civilian or military,
- developing a criterion for completeness of records (based on both personal identification items and items identifying death/disappearance of victims)
- marking the completeness of records in the database
- using records in some basic analysis.

Major results of the above-mentioned activities are discussed below.

1.4 ITEMS IDENTIFYING PERSONS, MISSING VALUES AND DUPLICATES OF RECORDS

In order to uniquely identify persons the following items should be available for every victim: the personal identification number (JMBG), names (first name(s), surname and father's name), DoB, PoB and PoR.[10] All these items were collected in the BBD project and are now contained in the BBD Database. As stated in Section 1, most likely they are incomplete (i.e. unavailable for some individuals) or contain errors (i.e. inconsistencies of reporting of one the same item among different persons; mainly related to data entry errors; sometimes errors result from reporting). Whereas nothing can be done about missing information, errors of data entry can be largely corrected (which the consultants have done to the extent it was possible without using additional sources of information). Below, in Tables 2 to 11, the completeness of (originally reported and cleaned) items is shown. Items included in these tables relate to the identification of individuals. We believe that as a minimum, all three names, a complete DoB and DoD[11] should be available in order to determine the identity of a victim. Without these items included for every person in the database, one cannot be sure who the victim was. This high standard cannot always be held in practice. So, if two names (first and family), year of birth and year of death are available in a record, the record can already be accepted as complete, subject to the requirement for this record to be unique (i.e. not duplicated). We used this practical criterion to distinguish between those cases that can be seen as complete, and thus of more value to users of the data, and those that have to remain less complete (of less value) at this stage. The relation of complete to less complete cases will most certainly evolve in the course of further improvements of the database. The record completeness (or value) criteria, and in particular marking missing values, are discussed in detail at the end of this section.

According to the above, Tables 2 to 11 consist, each, of three panels, "Complete", "Less Complete" and "All Cases". The panels "Complete" and "All Cases" are essential. The first

[10] DoB is date of birth, PoB place of birth, and PoR place of residence.
[11] DoD stands for date of death. In the BBD Database, the term "date of suffering" (DoSuff) is used for the same thing. So, the items called DoD and DoSuff are basically the same.

one shows statistics obtained from calculations made with the complete (active) records in the database (82,257), and the third one shows results based on all (active) records in the database (96,895). The fraction of the complete records in the database is about 85 percent. Statistics based on complete cases should be seen as minimum numbers which can be easily documented with a detailed personal record of victims and his/her death. Those based on all records are more deficient and therefore less reliable.

Table 2. Overview of the Completeness of JMBG

No of Digits in JMBG	Complete Cases		Less Complete Cases		Total All Cases	
	Number	Percent	Number	Percent	Number	Percent
13 Digits	35,250	42.9	2,122	14.5	37,372	38.6
1-12 Digits	279	0.3	7	0.0	286	0.3
Missing	46,728	56.8	12,509	85.5	59,237	61.1
Total	82,257	100.0	14,638	100.0	96,895	100.0

The JMBG is unavailable for a considerable portion of the database records (about 61% of all cases). The remaining records have the JMBG included, almost all as a 13-digit complete number (38.6%). Its quality is occasionally questionable except for dates of birth (7 first digits of JMBG), which are usually available in whole or as the year of birth. Table 6 below indicates that there are 411 records, (about 1% of the available JMBGs), that have inconsistent dates of birth when compared with the DoBs individually reported. Some of these DoBs are inconsistent due to the erroneous reporting by informants/sources. Another problem is related to the duplicated JMBGs, sometimes reported for both spouses and children as exactly the same number. Some 557 records, (equivalent to about 275 pairs; about 0.7% of the available JMBGs), were identified as duplicates on JMBG (see also Table 12 below and the related discussion). All these records were carefully checked and about a half of them marked for exclusion as duplicates. Thus, this problem is of a very small scale and can be easily handled. However, mainly because of the many missing values, we cannot use JMBG as part of the regular identification criterion of victims.

The next item needed for the proper identification of persons is the surname. Surnames are available for all individuals in the BBD Database, although in a few cases the names are misspelled or incomplete.

Table 3. Overview of the Completeness of First Name

FstName	Complete	Less Complete	Total (No.)	Total (%)
Available	82,257	14,619	96,876	100.0
Not Available	0	19	19	0.0
Total (No.)	82,257	14,638	96,895	100.0
Total (%)	84.9	15.1	100.0	-

The first name is also almost always available. In Table 3 only 19 first names are missing. This is a very small problem and excluding these records as less complete does not affect the database size.

Table 4. Overview of the Completeness of Father's Name

FaName	Complete	Less Complete	Total	Total (%)
Available	80,088	8,612	88,700	91.5
Not Available	2,169	6,026	8,195	8.5
Total	82,257	14,638	96,895	100.0
Total (%)	84.9	15.1	100.0	-

The third name, father's name, is unavailable for 8,195 persons, (8.5% of all cases), and excluding all these records as less complete would cause a considerable reduction of the database size (8.5%; Table 4). Incomplete records sometimes are incomplete on several items, however. Therefore, records incomplete on father's name might also have other deficiencies. Among the complete records, only 2,169 of them do not include father's name (2.6% of all complete), which is fairly acceptable.

Table 5. Overview of the Completeness of Year of Birth

YoB(cl)	Complete	Less Complete	Total	Total (%)
Available	82,257	5,208	87,465	90.3
Not Available	0	9,430	9,430	9.7
Total	82,257	14,638	96,895	100.0
Total (%)	84.9	15.1	100.0	-

YoB as reported individually (by informants or other sources different than JMBG) is missing for 9,430 cases (9.7% of all records; Table 5). All these records must be marked as less complete at this stage, thus also less reliable. Noteworthy, Table 5 shows results based on a cleaned copy of YoB, in which years later than 1995 and some missing values were replaced by YoB taken from JMBGs (if the latter were reasonable). In addition to that, YoBs with an extra "9"entered mistakenly after "19", as in "1996", were replaced whenever possible with the correct year (here: with "1965"). The number of such correction was small, however. Only 40 records could be improved in this way.

All in all, the loss of records related to unavailable reports of YoB is 9.7% of cases in the BBD Database. The good thing is that all available reports of YoB (90.3% of all cases) are consistent in the context of other related data items, such as e.g. date of death.

Table 6 below shows a sample of comparing two YoB items available in the BBD Database. The first YoB was reported individually by sources (witnesses, family, friends, published lists etc.) and the second can be separated from the JMBG (7 first digits; hereafter YoB(JMBG)). A high number of records in the database have consistent YoBs. Out of the total of 37,600 YoBs reported in the JMBGs, some 37,153 are fully consistent with individually reported YoBs (i.e. about 99%). Some 411 records have inconsistent YoBs in the JMBGs (about 1% of 37,600). Some 36 records have YoB(JMBG) available whereas the individually reported

YoB is unavailable; all those YoB(JMBG)s were inserted into the cleaned copy of YoB (used for analysis here).

Table 6. Overview of the Consistency of Year of Birth Reporting: Individually Reported versus JMBG-based YoBs[12]

YearB	YearB(JMBG)	Number
1904	1943	1
1907	1970	1
1908	1958	1
1911	1942	1
1919	1918	1
1920	1919	1
1923	1919	1
1923	1921	1
1923	1926	1
1928	1922	1
1928	1974	1
1929	1928	1
1929	1965	1
1930	1931	1
1931	1932	2
1931	1937	1
1931	1956	1
1933	1938	2
----------	----------	----------
1999	1990	1
Total		411

The cleaned YoB was created on the basis of the assumption that the YoB individually reported is correct. However, in records where only JMBG-based YoB was available, this YoB was taken as complete. This improved a total of 36 records, of which 31 were eventually marked as complete using the final record completeness (or value) criterion developed in this project. Altogether, a total of 90.3% BBD records were considered as having YoB available.

Regarding the cleaning of dates, note that usual mistakes included cases where original dates were entered in wrong cells, i.e. some DoBs were entered as DoDs (and vice versa). These dates could be easily detected from comparisons of these two types of dates. Secondly, in the BBD the year of events (e.g. births and also deaths) was entered as two last digits only (e.g. YoB entered as "45" represents "1945", but "92" might represent both "1992" and "1892"). For persons born in the last years of the 19th century, the year of birth was then unclear. Some of these cases could be detected and corrected by comparing YoB with YoD. In some corrections the JMBG-based DoB was very helpful. Finally, an unpleasant data entry error was sometimes made in dates by inserting an extra "9" (e.g. "97" instead of "74"). These errors were corrected also by using JMBG-based DoB. In some cases the JMBG was not available, and then an approximation of YoB was used (e.g. for dates ending with "97", YoB was assumed 1975, for "93" YoB=1935, for "95" YoB=1955 etc.). The cleaning of DoBs was

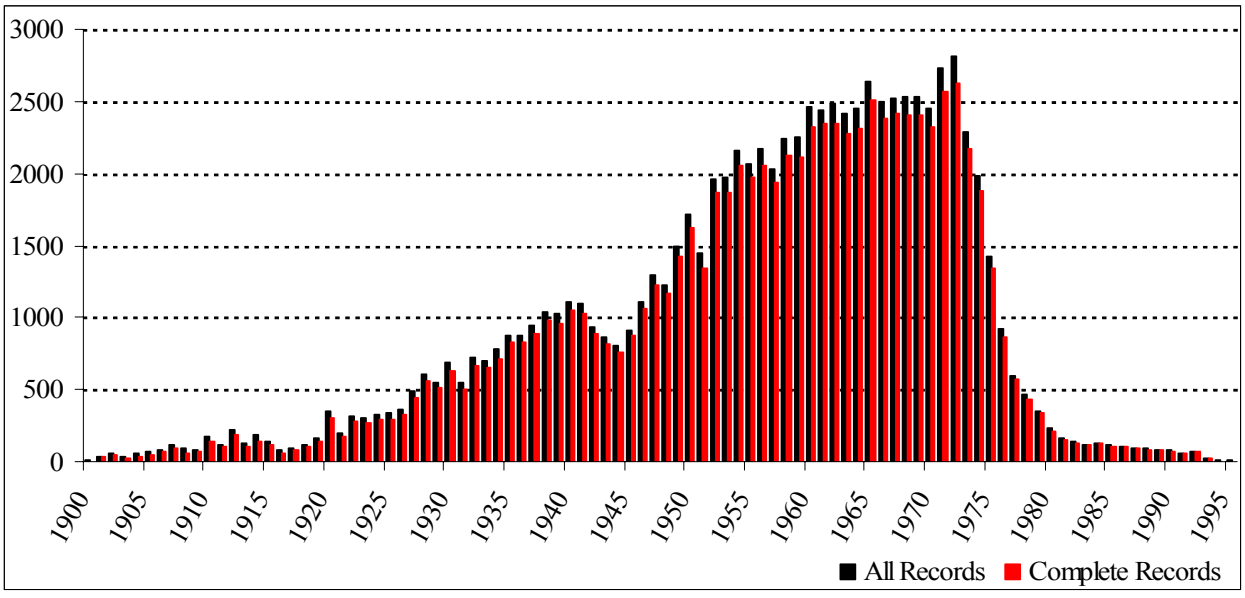
[12] Table 6 contains only a sample of all 411 records. Many records have been skipped for the ease of presentation. The term YoB is used interchangeably with YearB. These mean one the same thing: year of birth. See the list of abbreviations at the end of this report.

relatively successful which can be seen from the consistent age distribution of the victims. Table 7 below summarises the corrections on DoB. Eventually, 40 records were improved.

Table 7. Overview of the Corrections of Year of Birth (reported YoB vs. cleaned YoB (cl))

YearB	YearB (Cleaned)	Number
1993	1933	1
1993	1935	4
1993	1936	1
1993	1939	2
1993	1963	1
1994	1945	2
1994	1947	1
1994	1949	2
1995	1950	1
1995	1952	1
1995	1955	7
1995	1956	1
1995	1959	2
1996	1962	1
1996	1963	1
1996	1965	3
1996	1967	1
1997	1897	1
1997	1970	1
1997	1972	1
1997	1975	1
1998	1985	3
1999	1990	1
Total		40

Figure 1. BBD Records by Year of Birth (cl)



The next three charts (Figures 1 to 3) show the percentage distribution of, respectively, year, month and day of birth (all cleaned items, all records versus complete records). The purpose of presenting these charts is (a) to visualize possible outliers, (b) make sure that the selection of complete records does not change the distribution type of the original items.

Figure 2. BBD Records by Month of Birth (cl)

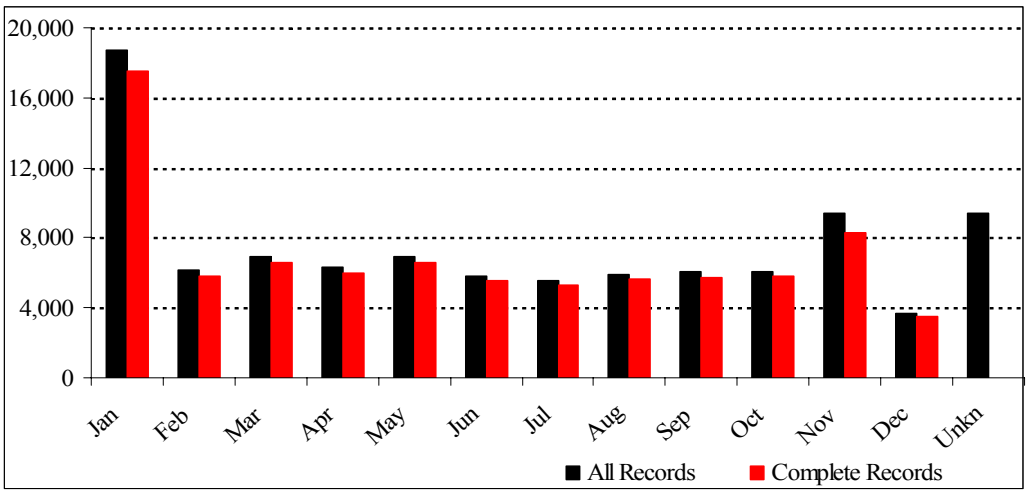
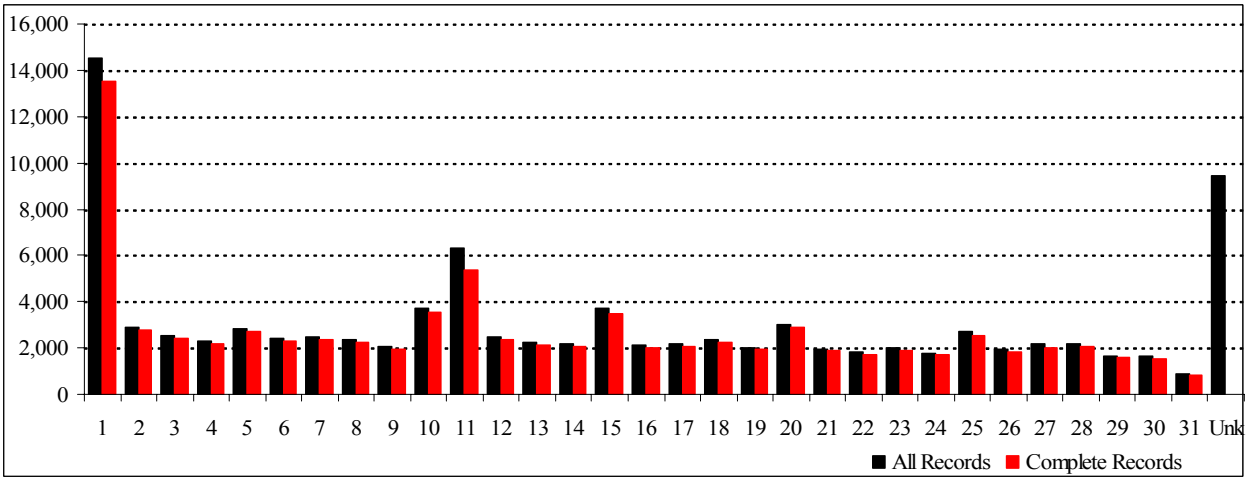


Figure 3. BBD Records by Day of Birth (cl)



A conclusion from Figures 1 to 3 is that no outliers are seen for YearB whereas there are clearly outliers in the charts of MonthB and DayB, and, secondly, complete records have in all three cases the same distribution type as the original items.

The outliers in MonthB are months of January (code "01") and November (code "11"). The outliers in DayB are again the 1st and 11th day. The codes of (especially) "01" and also "11" were obviously also used for cases with unclear or unknown month/day of birth. The effect of this misuse of complete codes is seen as exceptionally high levels of births in January (1st and 11th) and in November (again 1st and 11th). These days/months should be now considered less

complete. It would be much better to use codes, such as, for example, "99" or "Null" for unavailable days/months in dates.

One more issue is seen in the reporting of day of birth. "Round" days, such as 5th, 10th, 15th, 20th, and 25th, are clearly more frequently seen in Figure 3 than all other days. This problem, known in demography as age heaping, is present in the BBD too. The reason for it is that informants do not remember exact dates and instead they report the first "round" date close to the actual date of birth. This problem is a minor deficiency here and has no impact on the analytical results obtained from the database. It might have some impact though when using this data for court purposes; not all DoBs can be then taken as one hundred percent correct.

Data shown in Figures 1 to 3 are included in Tables 8 and 9. The data confirm the patterns displayed graphically in Figures 1 to 3. In these tables, next to the presentation of the usual three panels, i.e. "Complete", "Less complete" and "All Cases", a fourth one is introduced: "Drop Out Ratio" (hereafter: DO ratio). The DO ratio is a simple measure of the quantity of information lost when data on "Complete" cases are used instead of data on "All Cases". Drop-out ratios are substantially higher for the months of January and November than for all remaining months. The ratios are also relatively higher for the 1st and 11th days of each month. These high values confirm that the misuse of codes "01" and "11" is more common in generally less complete records which are rejected as less complete cases and excluded from further analysis.

Throughout the rest of this report, a frequent reference is made to the drop-out ratios which are considered to be measures of the quality of reporting and computerization of the BBD data for various groups of victims, such as, for example, municipalities (e.g. of birth or residence), ethnic or religious groups, or just the time and area of death.

Table 8. Overview of the Completeness of Month of Birth (MonthB)

MonthB	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Jan	17,486	21.3	1,225	8.4	18,711	19.3	6.5
Feb	5,799	7.0	334	2.3	6,133	6.3	5.4
Mar	6,548	8.0	346	2.4	6,894	7.1	5.0
Apr	6,020	7.3	286	2.0	6,306	6.5	4.5
May	6,564	8.0	353	2.4	6,917	7.1	5.1
Jun	5,524	6.7	292	2.0	5,816	6.0	5.0
Jul	5,315	6.5	283	1.9	5,598	5.8	5.1
Aug	5,648	6.9	275	1.9	5,923	6.1	4.6
Sep	5,762	7.0	296	2.0	6,058	6.3	4.9
Oct	5,783	7.0	260	1.8	6,043	6.2	4.3
Nov	8,273	10.1	1,106	7.6	9,379	9.7	11.8
Dec	3,508	4.3	181	1.2	3,689	3.8	4.9
Unkn	27	0.0	9,401	64.2	9,428	9.7	99.7
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 9. Overview of the Completeness of Day of Birth (DayB)

DayB	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
1	13,530	16.4	1,007	6.9	14,537	15.0	6.9
2	2,761	3.4	146	1.0	2,907	3.0	5.0
3	2,430	3.0	126	0.9	2,556	2.6	4.9
4	2,175	2.6	107	0.7	2,282	2.4	4.7
5	2,728	3.3	134	0.9	2,862	3.0	4.7
6	2,283	2.8	110	0.8	2,393	2.5	4.6
7	2,347	2.9	112	0.8	2,459	2.5	4.6
8	2,252	2.7	127	0.9	2,379	2.5	5.3
9	1,949	2.4	107	0.7	2,056	2.1	5.2
10	3,545	4.3	194	1.3	3,739	3.9	5.2
11	5,370	6.5	965	6.6	6,335	6.5	15.2
12	2,373	2.9	110	0.8	2,483	2.6	4.4
13	2,121	2.6	110	0.8	2,231	2.3	4.9
14	2,077	2.5	99	0.7	2,176	2.2	4.5
15	3,488	4.2	211	1.4	3,699	3.8	5.7
16	2,032	2.5	99	0.7	2,131	2.2	4.6
17	2,070	2.5	108	0.7	2,178	2.2	5.0
18	2,221	2.7	125	0.9	2,346	2.4	5.3
19	1,920	2.3	84	0.6	2,004	2.1	4.2
20	2,883	3.5	147	1.0	3,030	3.1	4.9
21	1,864	2.3	87	0.6	1,951	2.0	4.5
22	1,721	2.1	84	0.6	1,805	1.9	4.7
23	1,885	2.3	106	0.7	1,991	2.1	5.3
24	1,723	2.1	71	0.5	1,794	1.9	4.0
25	2,565	3.1	129	0.9	2,694	2.8	4.8
26	1,844	2.2	95	0.6	1,939	2.0	4.9
27	2,032	2.5	125	0.9	2,157	2.2	5.8
28	2,089	2.5	108	0.7	2,197	2.3	4.9
29	1,565	1.9	89	0.6	1,654	1.7	5.4
30	1,561	1.9	70	0.5	1,631	1.7	4.3
31	826	1.0	45	0.3	871	0.9	5.2
Unk	27	0.0	9,401	64.2	9,428	9.7	99.7
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Tables 8 and 9 contain specific statistics showing the specific levels of less complete dates (January 1st and 11th , November 1st and 11th). Records with these dates have not been excluded, at this stage, from the analysis. The details of DoBs are relevant, however, when comparing records from BBD with other sources (i.e. matching). Thus, for the purpose of matching the less complete details of DoBs should be replaced with the "99" or "Null" code.

Table 10. Overview of the Completeness of Place of Birth (PoB)

PoB	Complete Cases		Less Complete Cases		Total All Cases		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Known	67,784	82.4	5,308	36.3	73,092	75.4	7.3
Unknown	14,473	17.6	9,330	63.7	23,803	24.6	39.2
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 10 gives an overview of the availability of place of birth (PoB). About a quarter of records do not contain this information (24.6%). The reported PoBs (75.4%) include names of villages, settlements, towns, municipalities, as well as streets and other populated places. This

implies that PoB cannot be used as part of standard completeness or matching criteria. When studying individual records, (e.g. in manual checks of matching results), PoB might be helpful, though, to conclude definite matches or duplicates.

Place of residence is expressed in the BBD Database as a municipality (MoR; Table 11). The data entry was done by using numeric codes and a code book was provided for decoding (i.e. linking codes with names). Codes (and names) are available for almost all records in the database (only 765 records (0.8%) are empty on MoR). The MoR is one of the most complete and consistent items in the database.

Table 11. Municipality of Residence: An Overview of Selected Municipalities

No.	Municipality of Residence	Complete Records		Less Complete Records		Total All Records		Drop Out
		Number	Percent	Number	Percent	Number	Percent	Ratio
1	***SARAJEVO-TOTAL***	***12,156***	***14.8***	***1,806***	***12.3***	***13,962***	***14.4***	***12.9***
1	*1.1 CENTAR*	*1,477*	*1.8*	*105*	*0.7*	*1,582*	*1.6*	*6.6*
1	*1.2 HADZICI*	*601*	*0.7*	*55*	*0.4*	*656*	*0.7*	*8.4*
1	*1.3 ILIDZA*	*1,537*	*1.9*	*147*	*1.0*	*1,684*	*1.7*	*8.7*
1	*1.4 ILIJAS*	*479*	*0.6*	*77*	*0.5*	*556*	*0.6*	*13.8*
1	*1.5 NOVI GRAD*	*3,019*	*3.7*	*250*	*1.7*	*3,269*	*3.4*	*7.6*
1	*1.6 NOVO SARAJEVO*	*1,715*	*2.1*	*162*	*1.1*	*1,877*	*1.9*	*8.6*
1	*1.7 PALE*	*436*	*0.5*	*42*	*0.3*	*478*	*0.5*	*8.8*
1	*1.8 STARI GRAD*	*1,103*	*1.3*	*83*	*0.6*	*1,186*	*1.2*	*7.0*
1	*1.9 TRNOVO*	*312*	*0.4*	*38*	*0.3*	*350*	*0.4*	*10.9*
1	*1.10 VOGOSCA*	*568*	*0.7*	*139*	*0.9*	*707*	*0.7*	*19.7*
1	*1.11 UNSPECIFIED*	*909*	*1.1*	*708*	*4.8*	*1,617*	*1.7*	*43.8*
1	**SARAJEVO-TOTAL**	**12,156**	14.8	**1,806**	12.3	**13,962**	14.4	**12.9**
2	SREBRENICA	6,967	8.5	624	4.3	7,591	7.8	8.2
3	PRIJEDOR	3,976	4.8	1,309	8.9	5,285	5.5	24.8
4	ZVORNIK	3,520	4.3	593	4.1	4,113	4.2	14.4
5	BRATUNAC	3,081	3.7	359	2.5	3,440	3.6	10.4
6	VLASENICA	2,204	2.7	543	3.7	2,747	2.8	19.8
7	FOCA	2,188	2.7	510	3.5	2,698	2.8	18.9
8	MOSTAR	2,136	2.6	363	2.5	2,499	2.6	14.5
9	DOBOJ	1,995	2.4	267	1.8	2,262	2.3	11.8
10	ROGATICA	1,429	1.7	583	4.0	2,012	2.1	29.0
11	BANJA LUKA	1,399	1.7	426	2.9	1,825	1.9	23.3
12	BRCKO	1,430	1.7	237	1.6	1,667	1.7	14.2
13	VISEGRAD	1,322	1.6	331	2.3	1,653	1.7	20.0
14	GORAZDE	1,261	1.5	312	2.1	1,573	1.6	19.8
15	ZENICA	1,432	1.7	118	0.8	1,550	1.6	7.6
16	**TOTAL TOP 15**	**46,496**	56.5	**8,381**	57.3	**54,877**	56.6	**15.3**
17	REMAINING MUN.	35,574	43.2	5,678	38.8	41,252	42.6	13.8
18	OTHER-CROATIA	1	0.0	0	0.0	1	0.0	0.0
19	OTHER-UNKNOWN	186	0.2	579	4.0	765	0.8	75.7
20	**OVERALL TOTAL**	**82,257**	100.0	**14,638**	100.0	**96,895**	100.0	**15.1**

Only one record reports MoR from outside Bosnia (Other-Croatia; Table 11). This place might be a temporary residence of a BH citizen. Otherwise it would have to be excluded from analyses relating to the original population of BH.

Duplicates pose a serious problem in almost every database on war victims. In the BBD Database, duplicate control was conducted electronically at the stage of data entry by the authors of the BBD and was likely to be strong. This is demonstrated by the high number of duplicates among the records not marked as active (149, 841). Given the checking of

duplicates by the RCD, we did not expect to find a lot of duplicates among the active cases. Nevertheless, we had run additional checks for duplicates for the active cases as well; we did it two times in our project. First, duplicate searches were run when we studied data problems and about 4,400 records were then checked manually for the presence of duplicates. The purpose of these checks was to obtain an initial impression of the seriousness of the issue. Second, a systematic investigation of duplications was conducted at the stage of studying the overall coverage of the database (presented in Part III).

As a principle, one of the most comprehensive criteria for duplicate search should include the following items: first name, fathers name, surname, JMBG, DoB, PoB, and PoR. However, some of these items are inconsistently reported (e.g. PoB), many records have missing values on several of these items (father's name, DoB, JMBG), and some items contain errors (DoB). For these reasons, the procedure applied in duplicate checks should concentrate on a few of the available and well reported items, such as, for example, names and cleaned year of birth. Note that the availability of YoB is a good proxy for the availability of the entire DoB and that using the cleaned YoB helps avoiding the reporting bias in DoB. All records found identical on these items need to be checked manually, and some of them marked for retaining, some for deleting, and some as undecided. Note that in manual checks many more data items are usually compared than only those items used for selecting potential duplicates. Manual checks are therefore an essential step in assessing which records are duplicated and which are not.

We used three alternative criteria for selecting potential duplicates (comp. Table 12). Each subsequent criterion was run over not yet checked records. Although the searches cover major sources for duplication, they should not be seen as exhaustive, but as an attempt to investigate the seriousness of duplicates in the database. More duplicates were identified by subsequent additional attempts we made later.

Table 12. Overview of Duplicates by Duplicate-Find Query

Query 1:	Identity of the 3 first letters of all names (first, surname, father's name), year of birth (cleaned), and municipality of residence. From the 96, 895 records, a total of 1,095 records were selected as potential duplicates, all of them checked manually by comparing the following items: first name, surname, father's name (all in full), YoB(cl) and reported DoB, YoD(cl) and reported DoD, municipality of residence and of death. Some 489 records were marked as duplicates and the associated 486 records were not. Also the remaining 120 records were not duplicated.
Query 2:	Identity of JMBG (cleaned); no names or any other items included. All selected records (557) were checked manually on first name, surname, father's name (all in full), JMBG, DoB, DoD, municipality of residence and of death. Some 138 records (69 pairs) were found not to be duplicates, even though they had the same JMBG. Out of the remaining 419 records, some 210 appeared to be

	duplicates and 209 were not.
Query 3:	Identity of first name initial (1), surname initials (3), father's initial (1), cleaned year of birth, cleaned YoD. Manual checks of all selected records (2,741 selected) on: first name (in full), surname (in full), father's name (in full), JMBG, reported DoB, reported DoD. Some 361 records were marked as duplicates and the associated 359 as not duplicated. The remaining 2,021 records contained no associated pairs.

Altogether 4,393 records were selected as potential duplicates and were all checked manually. In total, 2,279 records were concluded as unrelated (i.e. not pairs), and the remaining 2,114 records were related pairs. Out of 2,114 associated records, some 1,054 were marked as non-duplicates, and 1,060 as duplicates. Duplicates account for 1.1% of all BBD records, and thus the scale of this problem is very minor.

Query 2, requiring the identity of JMBG for a record to be selected; names and other personal details were not part of the search criterion; made it possible to see all records with the same JMBG but differently spelled names, differences in DoBs and/or DoDs. Some records were obviously related to different persons even though they had the same JMBG. There were 138 such records (about 69 pairs). A half of these records should be considered as possibly having wrong JMBGs.

Table 13. Overview of BBD Records by the Criteria of Final Completeness

Record Completeness	Duplicate to Exclude	Availability of FamName	Availability of FstName	Availability of YoB*	Availability of YoD*	Number of Records
Complete	No	Yes	Yes	Yes	Yes	82,257
Less Complete	Yes	Yes	Yes	Yes	Yes	973
Less Complete	Yes	Yes	Yes	Yes	No	87
Less Complete	No	Yes	Yes	Yes	No	4,139
Less Complete	No	Yes	Yes	No	Yes	6,223
Less Complete	No	Yes	Yes	No	No	3,197
Less Complete	No	Yes	No	Yes	Yes	7
Less Complete	No	Yes	No	Yes	No	2
Less Complete	No	Yes	No	No	Yes	7
Less Complete	No	Yes	No	No	No	3
Total Less Complete	-	-	-	-	-	14,638

Checks of completeness and deficiencies of items identifying persons and duplicate checks were conducted in order to use their results in developing the criteria for final completeness of records in the database. It was decided to establish the final completeness based on five data items (see Table 13):

- duplicate
- surname
- first name

- year of birth (cleaned; YoB)
- year of death (cleaned; YoD)

Records marked as non-duplicates and with complete values on surname, first name, YoB and YoD were decided to be finally complete. All other records characterised by one or more deficiencies were marked as less complete.

Table 14. Overview of the Final Completeness of Records in BBD

Final Completeness	Number	Percent
Complete	82,257	84.9
Less Complete	14,638	15.1
Total	96,895	100.0

Table 13 and 14 point out that there are 82,257 records marked as finally complete (about 85% of all active records). These records can be safely used in research and analysis. The remaining 14,638 records (15%) are not marked as complete, i.e. at this stage they should be seen as less complete, due to one or more deficiencies. Note that among 14,638 deficient records, 11,342 records (about 77.5%) had just one deficiency, and 3,296[13] records (22.5%) two or more deficiencies (Table 15). Thus, a majority of less complete records were marked as such because of one shortcoming.

Table 15. Deficiencies of Records in BBD by Type and Frequency

Type of Deficiencies	Number of Deficiencies	Distribution of Deficiencies	
		One Deficiency	Two or More Deficiencies
1. Duplicate	1,060	973	87
2. Unknown SurName	0	0	0
3. Unknown FstName	19	7	12
4. Unknown YoB	9,430	6,223	3,207
5. Unknown YoD	7,428	4,139	3,289
Total (Number)*	17,937	11,342	6,595
Total (Percent)	100.0	63.2	36.8

Note:
**The total of 17,937 gives deficiencies; 14,638 relates to deficient records.*
*The number of deficient records **is not the same** as the number of deficiencies in these records.*

Table 15 further indicates that the most frequent deficiency of BBD records is the missing year (and date) of birth (9,430), the second most frequent is the missing year (and date) of death (7,428), and the third most prominent deficiency are the duplicates (1,060). These three shortcomings together cause the vast majority of record rejections in the final completeness criterion.

Summing up, it seems that the final completeness criterion proposed in this report efficiently eliminates records that are too incomplete to be reliably used in research and analysis.

[13] The number of 3.296 was obtained by subtracting the number of records with one deficiency (i.e. 11,342; see Table 15) from the overall total of deficient records in the BBD database (i.e. 14,638).

1.5 BASIC DEMOGRAPHIC CHARACTERISTICS

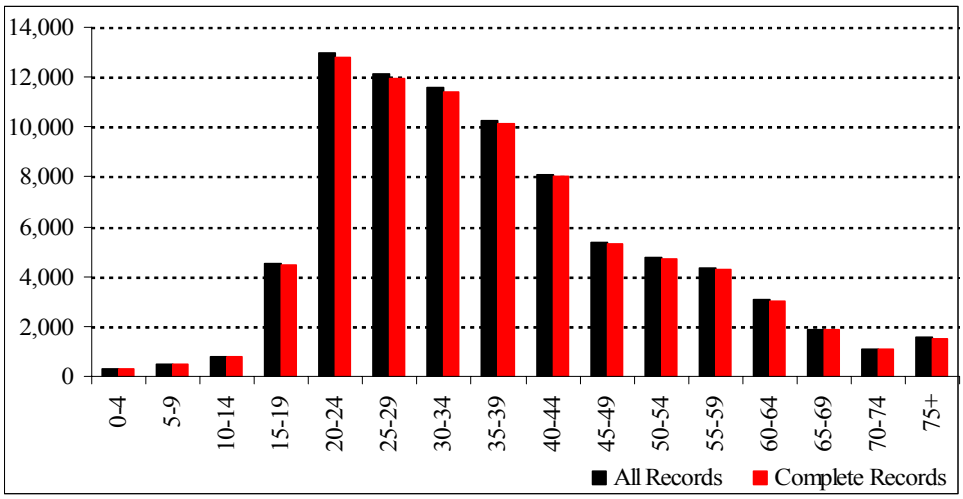
In this section, demographic characteristics of BBD victims are reviewed. They include age, sex, ethnicity, religion, and civilian-military status. In addition to these items, at the end of this section some attention is paid to the sources of BBD entries.

Age at death is one of the basic characteristics of every deceased. Age was not explicitly reported, however, in the BBD Database. Even though not reported, age can be easily calculated on the basis of the complete dates of birth and death, or on the basis of years of birth and death alone. Table 16 shows the age distribution of BBD victims, by five –year intervals, for the age obtained as a difference between the cleaned year of death and cleaned year of birth. A significant number of records have missing values on these two years (13,658; 14.1%) and age cannot be calculated. For the remaining records, where both years are available and complete, age was obtained as the above-mentioned difference and resulted in a fully consistent age distribution of the victims (Figure 4).

Table 16. Age at Death Distribution of BBD Victims

Age at Death	Complete Records		Less Complete Records		Total All Records	
	Number	Percent	Number	Percent	Number	Percent
0-4	308	0.4	9	0.1	317	0.3
5-9	494	0.6	1	0.0	495	0.5
10-14	762	0.9	10	0.1	772	0.8
15-19	4,451	5.4	65	0.4	4,516	4.7
20-24	12,797	15.6	181	1.2	12,978	13.4
25-29	11,964	14.5	154	1.1	12,118	12.5
30-34	11,417	13.9	145	1.0	11,562	11.9
35-39	10,160	12.4	114	0.8	10,274	10.6
40-44	7,998	9.7	104	0.7	8,102	8.4
45-49	5,323	6.5	59	0.4	5,382	5.6
50-54	4,714	5.7	54	0.4	4,768	4.9
55-59	4,310	5.2	34	0.2	4,344	4.5
60-64	3,042	3.7	30	0.2	3,072	3.2
65-69	1,888	2.3	9	0.1	1,897	2.0
70-74	1,092	1.3	8	0.1	1,100	1.1
75+	1,537	1.9	3	0.0	1,540	1.6
Unknown	0	0.0	13,658	93.3	13,658	14.1
Total	82,257	100.0	14,638	100.0	96,895	100.0

Figure 4. Victims Reported in BBD by Age at Death, Five-Year Intervals, Absolute Numbers



Most victims died at age 20 to 44 years; considerable numbers of deaths are also seen for age intervals 15 to 19 and 45 to 64 years. Very few deaths are observed for ages below 15 and beyond 64 years. This type of age distribution as shown in Table 16 and Figure 4 is typical for mortality from violent causes of death, different than the causes of natural mortality, (old age, diseases, congenital malformations etc.), and it is consistent with the age distribution expected for combatants engaged in a violent conflict. Noteworthy, the BDD also contains civilian death records, which we discuss below.

Figure 5. Male Victims Reported in BBD by Age, Five-Year Intervals, Absolute Numbers

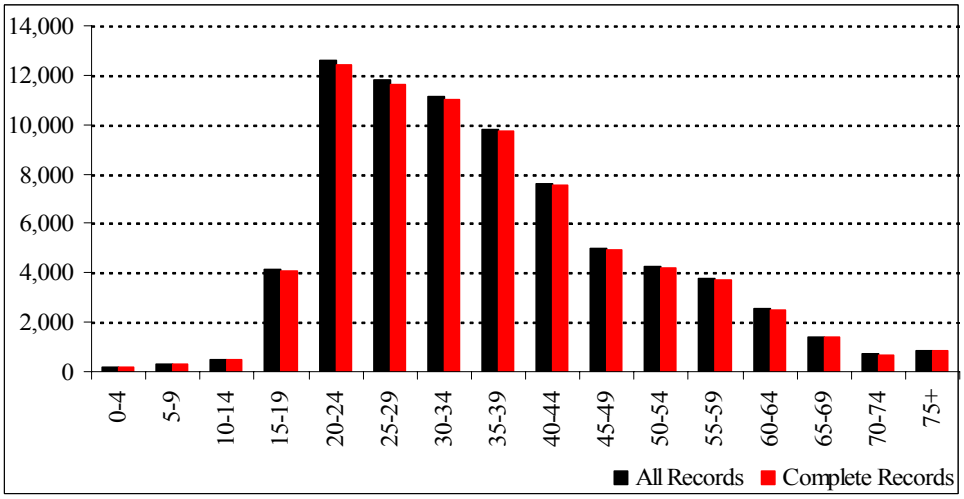
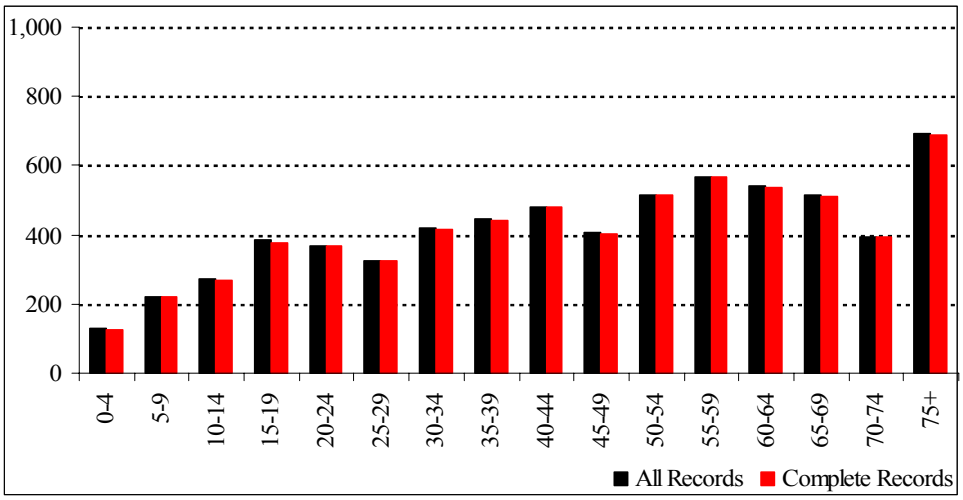


Figure 6. Female Victims Reported in BBD by Age, Five-Year Intervals, Absolute Numbers



Noteworthy, the age distribution for both sexes jointly (Figure 4) and the one of men (Figure 5) are almost identical, whereas the age distribution of women (Figure 6) differs considerably from that of men. Not only the scale of male mortality was much higher than that of females (up to maximally 13,000 male deaths versus up to 700 female deaths in one age interval), but also the age patterns are very distinctive in both cases. The age pattern of female deaths is not typical of combatant victims of a violent conflict neither does it resemble deaths from natural causes. There is little variation in their age-at-death pattern; women died relatively uniformly at all ages, which is more in line with a lack of a specific selection process with respect to age during the war operation.

All in all, the BBD records are mainly of men, and only marginally of women. The age pattern of male deaths strongly supports the violent character of their mortality, which most likely occurred in combat, whereas that of women indicates a low level process distributed uniformly over age.

Table 17. Victims Reported in BBD by Sex

Sex	Number	
@	1	0.0
1	3	0.0
B	1	0.0
D	3	0.0
M	87,505	90.3
Z	1	0.0
Ž	9,373	9.7
Unknown	8	0.0
Total	96,895	100.0

Note that the sex distribution used in the analysis of age patterns was based on the cleaned sex item. The originally reported sex is shown in Table 17. Except of some data entry errors, Table 17 indicates that a few of records were coded as “D” (3). These were records of

children ("dete" in B/C/S) whose sex was not reported as a man or woman. During the data cleaning, all records of children, records containing errors or missing values were recoded (based on the first names reported in these records) into two valid sex values of a man (M) or woman (Z).

Table 18. Victims Reported in BBD by Sex: Re-coded Cleaned Numbers

Sex	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Man	75,625	91.9	11,890	81.2	87,515	90.3	13.6
Woman	6,632	8.1	2,748	18.8	9,380	9.7	29.3
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 18 contains the sex distribution of the BBD records estimated according to the cleaned sex item. This distribution confirms that the number of female victims reported in BBD was indeed marginal (9.7% of all). Men are a vast majority of the BBD victims (90.3%). Missing values do not exist. The drop-out ratio is much higher for women than for men. A higher percentage of records of women have deficiencies compared to male records and thus need improvement.

Table 19. Ethnicity of Victims Reported in BBD

Ethnicity	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Croat	6,439	7.8	1,159	7.9	7,598	7.8	15.3
Muslim	53,878	65.5	10,125	69.2	64,003	66.1	15.8
Serb	21,679	26.4	3,147	21.5	24,826	25.6	12.7
Other	75	0.1	16	0.1	91	0.1	17.6
Unknown	186	0.2	191	1.3	377	0.4	50.7
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 19 contains the ethnic composition of BBD records. Ethnicity is a well-reported item, available for as many as 99.6% of all records (96,518 out of 96,895). For complete records (82,257) the availability of ethnicity is even higher (99.8%). Note as well that the final completeness criterion applied to exclude deficient and/or unreliable records from the analysis causes comparable losses of information among the ethnic groups. For Bosniacs about 15.8% of records are excluded as less complete, for Croats and Serbs 15.3% and 12.7 per cent, respectively. The reporting quality does not vary too much among the ethnic groups, the weakest being for Others (drop-out of 17.6%). This suggests an ethnicity-related bias is most likely not present in the database. Thus, in the case of BBD based statistics (complete records), one can speak with confidence about minimum numbers of war related deaths for every ethnic group.

The two following tables (21 and 22) are somewhat related to reporting of ethnicity. These tables confirm that also religion is a relatively well reported item; complete, consistent, and not requiring cleaning. Religion is obviously consistent with the ethnic distribution of victims; there is a high correspondence between religion and ethnicity.

Table 21. Religion of Victims Reported in BBD

Religion	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Catholic	6,438	7.8	1,149	7.8	7,587	7.8	15.1
Islam	53,917	65.5	10,127	69.2	64,044	66.1	15.8
Orthodox	21,667	26.3	3,140	21.5	24,807	25.6	12.7
Other	215	0.3	205	1.4	420	0.4	48.8
Unknown	20	0.0	17	0.1	37	0.0	45.9
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 22. Victims Reported in BBD by Ethnicity and Religion

Ethnicity	Catholic	Islam	Orthodox	Other	Unknown	Total
Croat	7,572	13	2	9	2	7,598
Muslim	1	63,951	2	21	28	64,003
Serb	7	4	24,796	16	3	24,826
Other	7	67	5	11	1	91
Unknown	0	9	2	363	3	377
Total	7,587	64,044	24,807	420	37	96,895

The next potentially important item in the BBD is the civilian-military status of the victims (hereafter "Status in War"). This item basically says how many victims were civilians and how many were militaries (Tables 23a and 23b).

Table 23a. Victims Reported in BBD by Status in War

Status in War	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Civilian	29,745	36.2	9,454	64.6	39,199	40.5	24.1
Policeman	989	1.2	45	0.3	1,034	1.1	4.4
Soldier	51,523	62.6	5,139	35.1	56,662	58.5	9.1
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 23b. Victims Reported in BBD by Status in War and Military Formation

Status in War	None	APZB	ARBIH	HV-HVO	JNA-VRS	**Total-No.**	Total-%
Civilian	39,191	0	4	0	4	**39,199**	40.5
Policeman	1	0	270	102	661	**1,034**	1.1
Soldier	19	549	30,485	5,609	20,000	**56,662**	58.5
Total-No.	**39,211**	**549**	**30,759**	**5,711**	**20,665**	**96,895**	100.0
Total-%	40.5	0.6	31.7	5.9	21.3	100.0	-

It seems that "Status in War" was defined in the BBD Database based on whether or not a deceased person was listed at the time of his/her death as a member of a military formation. In the Database, the sources for military formations are the lists of fallen soldiers and other military personnel published by the respective governments, armies and ministries of defence, as well as reports by individual informants. Among "All Cases" there are 39,199 civilians reported, 56,662 soldiers and 1,034 policemen (Table 23a). Compared with sources from the Office of the Prosecutor (ICTY), the number of soldiers seems high. For almost all militaries

(soldiers and policemen), their membership in a military formation is available in Table 23b (for all but 20), whereas almost all civilians remain unrelated to any of the military formations (except of 8 persons; 4 listed with Army of BiH and 4 with JNA forces). This high degree of consistency between these two items confirms that being a part of a military formation was a key used to distinguish between civilians and non-civilians in the BBD Database.

Note as well that the drop-out ratio reported in Table 23a is rather high for civilians (24.1%) as compared with soldiers (9.1%) and policemen (4.4%), suggesting that the quality of data on civilians was lower than on soldiers and implying that the distribution of "Status in War" is rather different when obtained from "Complete Records" versus from "All Records". In the former it is like 36 to 64 percent and in the latter case like 40 to 60 percent (civilians to militaries). The "Complete Records" distribution is most certainly based on underrepresented records of civilians, which urgently require an additional improvement in order to be reliably used in analysis.

The above item has something to do with the population losses in combat as opposed to losses in non-combat situations. However, even though "Status in War" is an approximation of this relationship, it does not offer a one-to-one relationship between the civilian and non-combat deaths, as well as the military and combat deaths, as not all civilians died in non-combat circumstances, and not all militaries in combat situations.

"Status in War" says also nothing about the actual circumstances of death and therefore, on its own, cannot be used for measuring numbers of victims of war crimes and crimes against humanity. Reporting a victim as a civilian (of whom we would even know that he/she died in a combat situation), does not yet guarantee that this person can be taken as a legitimate victim of a violation of the International Humanitarian Law, even if his/her death was caused by a conflict-related factor, such as, the above-mentioned combat situation. The civilian/military status alone cannot be used in measuring collateral losses among civilians going beyond civilian losses expected as proportional in a military confrontation. Moreover, also civilians occasionally actively participated in combat and were killed in combat but were reported as civilians in databases such as the BBD. The opposite would be of course true for soldiers who often were killed in non-combat situations, executed, died as prisoners of war, or of conflict-induced diseases etc., but because of their official status were reported as militaries in the sources published by the respective armies.

Despite of all these shortcomings, this particular item, i.e. "Status in War" might be still invaluable in assessment of the overall character of the 1992-95 conflict in Bosnia. In particular, it could answer the question about the extent of civilian lives lost in the context of incidents where the armies and other armed forces were engaged in achieving their military objectives. The analysis of the civilian death, in particular in combination with date and place of death in the BDD can also increase our insight in killings that occurred in places where there was no military confrontation at that moment. So, improving the reporting of civilians in the BBD is a high priority.

The authors of the BBD tried to compensate for the above-mentioned deficiencies of "Status in War" by providing an additional item related to combat versus non-combat deaths. (Based

on the original item: Suffering-level 2). Table 24 below gives a short overview of reporting of "Combat vs. Non-Combat Deaths" in the BBD. A very high percent of original "active" records have missing values on this item (96.4%). Among the complete records the fraction is even higher. So, as of now, this item is of no use at all.

Table 24. Combat versus Non-Combat Deaths of Victims Reported in BBD

Status At Death	Total-No.	Total-%
Non-combat activities	1,297	1.3
Prisoner of war	2,178	2.2
Unknown/Unavailable	93,420	96.4
Total	96,895	100.0

Finally, some information is available in the BBD Database about the military forces that caused deaths of civilians and soldiers. Table 25 give an overview of these forces (for "All Cases" only). For about 85% of victims the forces that brought death are unavailable implying that using this item in analysis makes at this stage no sense at all.

Table 25a. Victims Reported in BBD by Status in War and Cause of Death - Military Formation

Status in War	None	APZB	ARBIH	HV-HVO	JNA-VRS	**Total-No.**	Total-%
Civilian	33323	159	185	12	5520	**39,199**	40.5
Policeman	876	7	61	1	89	**1,034**	1.1
Soldier	47862	709	2130	47	5914	**56,662**	58.5
Total-No.	**82,061**	**875**	**2,376**	**60**	**11,523**	**96,895**	100.0
Total-%	84.7	0.9	2.5	0.1	11.9	100.0	-

Finally, we close this section by presenting a brief overview of sources for records contained in the BBD database (Table 26).

Table 26. Overview of Sources Reported in BBD

Code of data origin	Number
A	4,277
E	5,453
M	10,660
N	18,376
P	558
S	864
V	17,326
Y	65
None	39,316
Total	96,895

From the document "List of Sources" it is clear that several big and reliable sources were used, such as the ICRC list of missing persons for Bosnia, the BH Commission for Tracing Missing Persons, and army records. Several sources were smaller and likely less reliable, for example the Prijedor Book of Missing. The largest portion of records were collected from (eye)

witness statements, grave tombs, press reports, government sources, non-governmental organizations etc. The version of the BBD analysed by the consultants has no specific variables listing the sources for each case. From our work in the RDC, we know that the RDC staff can trace each case back to the source material, but an outside person will not be able to do that.

1.6 TIMING AND LOCATION OF DEATHS/MISSING

In this section we shortly summarize findings related to items characterizing death or missing. Basically three items are discussed here: date (DoD), cause (CoD) and place (i.e. municipality; MoD) of death/disappearance. Except for the date of death/disappearance (hereafter: date of death), cause and place are well reported. A major problem with the date is its incompleteness and errors in reporting the year of death (YoD). Whereas nothing can be done to improve the availability of YoD, errors can be (and were) largely corrected by analysing the records and comparing YoD with YoB. Similar problems (i.e. data entry mistakes) were noted with reporting YoD as with YoB.

Table 27 shows the originally reported year of death. As many as 6,998 records do not have YoD available. These records are practically useless for the quantification of numbers of war victims. In addition to this, several records are reported with YoD from before 1991 (31), 444 in 1991, and another group of records have YoD from after 1995 (12). All these years were thoroughly studied one by one. Mistakes were identified and corrected, especially in the years from before 1991. It is clear that in cases of specific analyses related to a given time span and a given territory these records would be automatically excluded.

Table 27. BBD Records by the Reported Year of Death

YearD	Number	YearD	Number
Unknown	6,998	*Continued:*	
1901	1	1973	1
1903	2	1974	1
1904	4	1989	1
1905	4	1991	444
1906	1	1992	42,467
1909	4	1993	18,621
1919	2	1994	9,384
1949	1	1995	18,938
1952	1	1996	10
1959	2	1998	1
1969	4	1999	1
1972	2	Total	96,895

Table 28. BBD Records by the Cleaned Year of Death

YaerD	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
1992	37,740	45.9	4,752	32.5	42,492	43.9	11.2
1993	17,439	21.2	1,195	8.2	18,634	19.2	6.4
1994	8,970	10.9	422	2.9	9,392	9.7	4.5
1995	18,108	22.0	841	5.7	18,949	19.6	4.4
Unknown	0	0.0	7,428	50.7	7,428	7.7	100.0
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 28 summarizes the results of cleaning and improving the year of death. The cleaning was successful and many YoDs were improved. Note that among the complete records, no YoD is seen from before 1992. The completeness criterion was defined in the way which excluded these records from those considered as complete.

Table 29. BBD Records by Month of Death

MonthD	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Jan	4,705	5.7	1,094	7.5	5,799	6.0	18.9
Feb	2,234	2.7	143	1.0	2,377	2.5	6.0
Mar	2,552	3.1	112	0.8	2,664	2.7	4.2
Apr	4,490	5.5	443	3.0	4,933	5.1	9.0
May	8,654	10.5	941	6.4	9,595	9.9	9.8
Jun	12,446	15.1	1,481	10.1	13,927	14.4	10.6
Jul	19,046	23.2	1,305	8.9	20,351	21.0	6.4
Aug	7,261	8.8	621	4.2	7,882	8.1	7.9
Sep	7,100	8.6	480	3.3	7,580	7.8	6.3
Oct	5,425	6.6	422	2.9	5,847	6.0	7.2
Nov	4,198	5.1	328	2.2	4,526	4.7	7.2
Dec	4,126	5.0	292	2.0	4,418	4.6	6.6
Unkn	20	0.0	6,976	47.7	6,996	7.2	99.7
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Except for a large number of records with the missing month of death (6,996 or 7.2%), no other problems are seen with MoD (Table 29). The three outliers, May, June and July, in 1992 and 1995 are most probably related to the intensity of killing in the Autonomous Region of Krajina and Eastern border with Serbia in 1992 and in Srebrenica in 1995. Below we briefly check this hypothesis by reviewing years of death year by year in the 1992-95 period (Table 30).

Table 30 pinpoints that indeed the years 1992 and 1995 contributed to the high numbers of killings and disappearances in the months May (1992), June (1992) and July (1995). Some more outliers are seen in this table too: July-August 1992, January 1993, June-July 1993, January-February 1994, and November 1994. All these dates can be linked to specific war episodes discussed in historical reports on the 1992-95 conflict in Bosnia.

Table 30. Complete BBD Records by Year and Month of Death, 1992 to 1995

MonthD	1992		1993		1994		1995	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Jan	882	2.3	1,868	10.7	1,315	14.7	640	3.5
Feb	62	0.2	1,078	6.2	805	9.0	289	1.6
Mar	116	0.3	1,315	7.5	621	6.9	500	2.8
Apr	1,612	4.3	1,620	9.3	686	7.6	572	3.2
May	6,143	16.3	1,100	6.3	600	6.7	811	4.5
Jun	8,699	23.0	2,075	11.9	632	7.0	1,040	5.7
Jul	5,676	15.0	2,719	15.6	526	5.9	10,125	55.9
Aug	4,458	11.8	1,363	7.8	572	6.4	868	4.8
Sep	3,351	8.9	1,267	7.3	553	6.2	1,929	10.7
Oct	2,562	6.8	1,000	5.7	779	8.7	1,084	6.0
Nov	1,958	5.2	933	5.4	1,171	13.1	136	0.8
Dec	2,211	5.9	1,099	6.3	707	7.9	109	0.6
Unk	10	0.0	2	0.0	3	0.0	5	0.0
Total	37,740	100.0	17,439	100.0	8,970	100.0	18,108	100.0

Note: Highlighted are values >10% in a given year

Table 31. BBD Records by Day of Death

DayD	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
1	6,631	8.1	1,908	13.0	8,539	8.8	22.3
2	1,999	2.4	189	1.3	2,188	2.3	8.6
3	1,949	2.4	134	0.9	2,083	2.1	6.4
4	2,082	2.5	175	1.2	2,257	2.3	7.8
5	2,307	2.8	150	1.0	2,457	2.5	6.1
6	1,834	2.2	136	0.9	1,970	2.0	6.9
7	2,168	2.6	132	0.9	2,300	2.4	5.7
8	2,405	2.9	178	1.2	2,583	2.7	6.9
9	2,187	2.7	160	1.1	2,347	2.4	6.8
10	3,293	4.0	286	2.0	3,579	3.7	8.0
11	2,752	3.3	302	2.1	3,054	3.2	9.9
12	9,181	11.2	507	3.5	9,688	10.0	5.2
13	3,024	3.7	200	1.4	3,224	3.3	6.2
14	2,747	3.3	193	1.3	2,940	3.0	6.6
15	2,519	3.1	288	2.0	2,807	2.9	10.3
16	2,652	3.2	221	1.5	2,873	3.0	7.7
17	2,055	2.5	132	0.9	2,187	2.3	6.0
18	2,194	2.7	180	1.2	2,374	2.5	7.6
19	2,307	2.8	176	1.2	2,483	2.6	7.1
20	2,933	3.6	255	1.7	3,188	3.3	8.0
21	2,230	2.7	139	0.9	2,369	2.4	5.9
22	1,977	2.4	145	1.0	2,122	2.2	6.8
23	2,464	3.0	180	1.2	2,644	2.7	6.8
24	2,217	2.7	177	1.2	2,394	2.5	7.4
25	2,541	3.1	243	1.7	2,784	2.9	8.7
26	2,206	2.7	147	1.0	2,353	2.4	6.2
27	2,280	2.8	186	1.3	2,466	2.5	7.5
28	2,162	2.6	143	1.0	2,305	2.4	6.2
29	1,709	2.1	132	0.9	1,841	1.9	7.2
30	2,166	2.6	201	1.4	2,367	2.4	8.5
31	1,066	1.3	67	0.5	1,133	1.2	5.9
Unk	20	0.0	6,976	47.7	6,996	7.2	99.7
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

DayD	BiH	Srebrenica	Percent in Srebrenica
1	51	10	0.1
2	35	0	0.0
3	50	0	0.0
4	68	0	0.0
5	30	0	0.0
6	18	0	0.0
7	114	88	1.2
8	24	0	0.0
9	23	1	0.0
10	56	16	0.2
11	747	558	7.7
12	6,724	5,910	81.7
13	691	396	5.5
14	220	77	1.1
15	114	33	0.5
16	71	11	0.2
17	139	81	1.1
18	58	12	0.2
19	63	1	0.0
20	84	7	0.1
21	111	3	0.0
22	31	4	0.1
23	91	3	0.0
24	50	0	0.0
25	56	0	0.0
26	80	3	0.0
27	67	1	0.0
28	142	6	0.1
29	39	1	0.0
30	57	5	0.1
31	21	3	0.0
Total	10,125	7,230	100.0

Note: For Srebrenica - Complete records only

Table 31 gives an overview of the day of death reporting. The picture is similar to that already discussed for month of death. One clear single outlier is the day "12". As indicated in the associated "July 1995" part of Table 30, this day can be linked in about 90 % of cases to Srebrenica. It is striking, however, that all these victims are reported as killed or gone missing on the 12th July, not on the 11th.

Table 32. Cause of Death of Victims Reported in BBD

Cause of Death	Complete Records		Less Complete Records		Total All Records		Drop Out
	Number	Percent	Number	Percent	Number	Percent	Ratio
Butchering	244	0.3	284	1.9	528	0.5	53.8
Forced suicide	7	0.0	3	0.0	10	0.0	30.0
Granate	7,566	9.2	701	4.8	8,267	8.5	8.5
Human shield,	0	0.0	1	0.0	1	0.0	100.0
Killed	55,191	67.1	11,019	75.3	66,210	68.3	16.6
Maltreated, Ki	205	0.2	132	0.9	337	0.3	39.2
Mine	521	0.6	15	0.1	536	0.6	2.8
Missing	16,299	19.8	2,359	16.1	18,658	19.3	12.6
Poisoning	1	0.0	0	0.0	1	0.0	0.0
Sniper	2,222	2.7	121	0.8	2,343	2.4	5.2
Wounds - Gra	1	0.0	0	0.0	1	0.0	0.0
Unknown/Una	0	0.0	3	0.0	3	0.0	100.0
Total	82,257	100.0	14,638	100.0	96,895	100.0	15.1

Table 32 focuses on the cause of death. No major problems are seen; missing values are a small fraction of all entries.

Note that cause of death was a self-reported item in the BBD project. Table 32 contains a tabulation that is based on English translation of the originally reported categories; no further grouping of original causes was made. It is rather clear from Table 32 that some categories could be combined, for example:

- "maltreated/killed" with "wounds, maltreatment, killed",
- "wounds-granate" with "granate".

Note also that "missing" is listed as one of the causes of death, but in fact "missing" should be seen as yet unknown cause of death.

Finally, Table 33 below gives (a sample of) results related to the municipality of death. Only a small number of records have a missing value on MoD (1,828; 1.9%). Another 688 records have place of death from outside Bosnia and Herzegovina. These records will be automatically excluded from all analyses for any territory within Bosnia. Sarajevo, Srebrenica, Prijedor, Zvornik, and Mostar belong to the territories with the highest coverage in this database.

Table 33. Municipality of Death of Victims Reported in BBD

No.	Municipality of Suffering	Complete Records		Less Complete Records		Total All Records		Drop Out
		Number	Percent	Number	Percent	Number	Percent	Ratio
1	***SARAJEVO-TOTAL***	***12,704***	***15.4***	***1,669***	***11.4***	***14,373***	***14.8***	***11.6***
1	*1.1 CENTAR*	*1,492*	*1.8*	*96*	*0.7*	*1,588*	*1.6*	*6.0*
1	*1.2 HADZICI*	*661*	*0.8*	*60*	*0.4*	*721*	*0.7*	*8.3*
1	*1.3 ILIDZA*	*1,767*	*2.1*	*150*	*1.0*	*1,917*	*2.0*	*7.8*
1	*1.4 ILIJAS*	*842*	*1.0*	*84*	*0.6*	*926*	*1.0*	*9.1*
1	*1.5 NOVI GRAD*	*2,512*	*3.1*	*180*	*1.2*	*2,692*	*2.8*	*6.7*
1	*1.6 NOVO SARAJEVO*	*1,437*	*1.7*	*134*	*0.9*	*1,571*	*1.6*	*8.5*
1	*1.7 PALE*	*205*	*0.2*	*40*	*0.3*	*245*	*0.3*	*16.3*
1	*1.8 STARI GRAD*	*1,140*	*1.4*	*70*	*0.5*	*1,210*	*1.2*	*5.8*
1	*1.9 TRNOVO*	*924*	*1.1*	*37*	*0.3*	*961*	*1.0*	*3.9*
1	*1.10 VOGOSCA*	*627*	*0.8*	*129*	*0.9*	*756*	*0.8*	*17.1*
1	*1.11 UNSPECIFIED*	*1,097*	*1.3*	*689*	*4.7*	*1,786*	*1.8*	*38.6*
1	**SARAJEVO-TOTAL**	**12,704**	**15.4**	**1,669**	**11.4**	**14,373**	**14.8**	**11.6**
2	SREBRENICA	8,681	10.6	696	4.8	9,377	9.7	7.4
3	PRIJEDOR	3,554	4.3	1,238	8.5	4,792	4.9	25.8
4	ZVORNIK	3,383	4.1	576	3.9	3,959	4.1	14.5
5	MOSTAR	2,353	2.9	352	2.4	2,705	2.8	13.0
6	BRATUNAC	2,343	2.8	313	2.1	2,656	2.7	11.8
7	FOCA	2,061	2.5	491	3.4	2,552	2.6	19.2
8	BIHAC	2,143	2.6	190	1.3	2,333	2.4	8.1
9	DOBOJ	1,845	2.2	216	1.5	2,061	2.1	10.5
10	BRCKO	1,767	2.1	230	1.6	1,997	2.1	11.5
11	ROGATICA	1,319	1.6	560	3.8	1,879	1.9	29.8
12	VLASENICA	1,355	1.6	494	3.4	1,849	1.9	26.7
13	GORAZDE	1,440	1.8	326	2.2	1,766	1.8	18.5
14	TRAVNIK	1,507	1.8	161	1.1	1,668	1.7	9.7
15	TUZLA	1,263	1.5	400	2.7	1,663	1.7	24.1
16	**TOTAL TOP 15**	**47,718**	**58.0**	**7,912**	**54.1**	**55,630**	**57.4**	**14.2**
17	REMAINING MUN.	33,445	40.7	5,383	36.8	38,828	40.1	13.9
18	OTHER-BOSNIA-HERZ.	18	0.0	3	0.0	21	0.0	14.3
19	OTHER-CROATIA	148	0.2	266	1.8	414	0.4	64.3
20	OTHER-MONTENEGRO	39	0.0	2	0.0	41	0.0	4.9
21	OTHER-SERBIA	109	0.1	16	0.1	125	0.1	12.8
22	OTHER-SLOVENIA	6	0.0	2	0.0	8	0.0	25.0
23	OTHER-UNKNOWN	774	0.9	1,054	7.2	1,828	1.9	57.7
24	**OVERALL TOTAL**	**82,257**	**100.0**	**14,638**	**100.0**	**96,895**	**100.0**	**15.1**

1.7 CONCLUDING REMARKS

The consultancy team assessed data deficiencies in the first place, and to a limited extent also the biases resulting from the method of data collection and the quality of reporting. Our assessment is based on the July 2006 version of the BBD and included the following activities:

- reviewing all original items and inspection of errors,
- cleaning obvious deficiencies that could be cleaned without using additional sources or analysis of data (e.g. spelling errors or misplaced information),
- cleaning of deficiencies involving comparisons of items (i.e. checking logical links between items; e.g. comparisons of dates, such as DoB and DoD, or DoB with JMBG etc.),
- re-coding items using code books,

- duplicate checks and marking duplicated records for removal,
- checking availability of items needed to uniquely identify victims
- checking availability of items needed to uniquely identify the death or disappearance of victims and circumstances of these events,
- checking availability of items needed to uniquely identify victims' status in war (civilian versus military),
- developing a criterion for the final completeness of records (based on both personal identification items and items identifying death/disappearance of victims)
- marking the final completeness of records in the database

The above-mentioned activities were completed as planned and are discussed in detail in Sections 1.1 to 1.6 of Part I of the assessment report. Below, we generally summarize the main findings.

- The BBD database contains information that was collected mainly on the basis of self-reporting by informants, that provided this information voluntarily, or was taken from overall sources on war-related victims, such as press reports, books, missing persons lists, government or NGO sources. No documents were required to prove statements of the respondents. For these reasons there might be some inconsistent and less reliable records included there as well.
- Even though it is the largest existing database on Bosnian war victims, the BBD should not be used alone but whenever possible one should use it together with other sources on war victims. This will prevent from producing biased statistics and historically incorrect pictures.
- The BBD database may be approaching its limits. This claim is suggested by RDC's observation that including new cases brings only marginal improvements; this implies that most cases have already been placed in the database (the RDC meeting in Sarajevo, 30 June - 1 July 2006).
- It is a large database which contains 96,895 active (or approved) cases out of a total of 246,736 computerized records. We assessed both parts of it, but the part marked as "active" records by the authors of BBD was central to us in our work.
- A brief examination of the part not marked as "active" confirmed that many of these records were indeed too deficient to conclude them "approved" and most of them were duplications of the active records.
- For the active cases, two groups of items were inspected on the first place: personal identification items and event (i.e. death or disappearance) identification items.
- Only a few problems were encountered within these two groups.
- Both groups were only slightly affected by data entry errors, or misplaced information, and more impact was seen for missing values.
- Except for missing values, all other deficiencies (excluding duplicates – these were studied separately) can be seen as extremely minor; many of them can be easily corrected by studying the records in the database and/or checking in the original source material what actually is wrong.
- Missing values are not a database problem. Missing values is a reporting problem; these were the informants that were unable to provide certain pieces of information to the BBD developers which resulted in incompleteness of certain data items in the database.

- Because of the missing values, not all of the active records could be considered complete. About 85% of active records were declared complete (82,257 out of 96,895 active) and about 15% of active records were concluded less complete (14,638). The less complete records (of mainly civilians, women of ethnicity "Other") can and should be improved and their completeness revised accordingly.
- The criterion for the final completeness of cases was based on the availability of the following items: surname, first name, year of birth (cleaned; YoB), and year of death (cleaned; YoD).
- Records marked as non-duplicates and with valid values on surname, first name, YoB and YoD were decided to be finally complete. All other records characterised by one or more deficiencies were marked as less complete. [14]
- Note that in fact more items are required for a record to be complete (e.g. JMBG, PoB, PoR, PoD, and CoD). But an extended completeness criterion would cause even more rejections of BBD records, which seems unnecessary. The quantity of information availably in the BBD complete records implies that these records can be already easily confirmed by cross-referencing the BBD material with other sources on war-related deaths.
- The most frequent deficiency of BBD records is the missing year (and date) of birth (9,430 or 9.7%), the second most frequent is the missing year (and date) of death (7,428 or 7.7%), and the third most prominent deficiency are the duplicates (1,060 or 1.1%).[15] These three shortcomings together cause almost all record rejections in the final completeness criterion. In relative terms, their scale is small, however.
- About 77.5% of rejected records (11,342 out of 14,638) are characterised by having a single deficiency. About 22.5% of rejected records are deficient on two or more items (3,296). This confirms the observation that deficient records tend to have missing values on one dimension only which is easy to repair.
- Records marked as complete can be relatively safely used in analysis. Statistics obtained based on complete records are the minimum or "at least" numbers and can be documented by relatively complete data from the database. However, comparisons among categories may be biased due to uneven rates of under-reporting (see Part III for more discussion).
- Ethnicity is available for practically all records in the database (0.4% missing), thus the availability is not an issue. Moreover, the quality of reporting is rather uniform among the ethnic groups and practically no ethnicity-related bias is present.
- Regarding the civilian-military status, called in the database "Status in War", it is reported on the basis of official military lists and other relevant sources. The RDC staff has visited 366 military and civilian cemeteries, gathered books and photos from military memorials, and collected a number of other appropriate sources that were all used to decide about the status of individuals. "Status in War" was obviously strictly based on the available sources and no arbitrary decisions were made with respect to the coding.

[14] Place of death of almost all cases was within the territory of Bosnia and Herzegovina, and practically all victims were born in BiH as well.

[15] 1,060 is a minimum number of duplicates in the database. More duplications were found and discussed in this report too. Here the minimum is used in the context of the final completeness criterion. The declarations of cases complete or less complete will have to be updated by the authors of BBD after they will clean the duplicates.

- Thus, "Status in War" merely shows the reporting of victims in military versus civilian sources. As of now, about 40% of victims are reported as civilians and 60% as soldiers (including the policemen). Among the complete records, the respective fractions are 36 and 64 percent.
- During our BBD project and other studies related to the registration of victims of war it came to our attention that some victims reported as soldiers according to official military lists, would be as well claimed civilians in civilian sources, and vice-versa. In particular, some military records could have been created by authorities in response to applications from the relatives of the deceased for the post-mortal benefits after the deceased. Secondly, some families might have found it honourable to bury their deceased among the defenders in military cemeteries or to publish their names on defenders' lists, even if the actual circumstances of death were not necessarily directly related to combat. These practices likely lead to over-reporting of soldiers and under-reporting of civilians in the sources. In consequence of these and other similar practices, civilians are in our opinion underrepresented in "Status in War".
- The civilian victims are also underrepresented among the complete records in the BBD as the drop-out ratio[16] for civilians (24.1%) is much higher than for soldiers (9.1%) or policemen (4.4%). Relatively many more civilians have been marked as less complete records whereas records of militaries were generally more complete.
- Improving the records of civilians is now a high priority and the analysis of the BBD data by "Status in War" should be postponed until a better data on civilians will become available. Especially, the military records reported as well in civilian sources should be reviewed and possibly revised. Secondly, the completeness of those records of civilians that are now marked as less complete should be enhanced.
- It is important to emphasize that "Status in War" does not provide correct insights in relation to victims of combat versus non-combat situations. Neither does it inform about legitimate victims of violations of the International Humanitarian Law or the Law of War. "Status in War" is a simple measure of whether or not a person was a member of a military/police formation at the time of death, (or generally was a defender), or a civilian. As such it offers a good basis for a further more specific investigation into issues related to involvement of victims in combat or proportionality of civilian losses.
- Being aware of the above mentioned differences, the authors of the BBD have made some attempts to shed some light on combatants versus non-combatants issues (data item: Suffering-Level 2). At this stage, however, these attempts cannot be concluded successful as the fraction of missing values on "Suffering-Level 2" is about 96% which disqualifies using this item in any analysis at all.
- The same is true in relation to another item, "Cause of Death – Military Formation", intended to describe which military formation caused deaths of victims. Again the 85% of values on this item are missing and this prevents from using this item at this stage.
- Also the mass grave part of the BBD Database (about 2,217 victims reported as exhumed from mass graves), is not ready for use as largely incomplete. The overall

[16] The drop-out ratio is a simple measure of the relationship between the number of less complete records and the overall number of all records in a given category. The ratio shows the percentage of less complete records in the respective total.

total of human remains exhumed from the mass graves (5 or more sets of remains) in Bosnia and Herzegovina was reported as 10,790 at the end of 2005 by the Federal Commission for Tracing Missing Persons. The overall total of identified persons was lower, (due to commingled remains and not-yet-available identifications), and equalled 8,724 persons. These numbers do not include statistics from the RS authorities. At present, the statistics are much higher. The RDC team clarified to us that exhumations part has a supportive role in then database and is not meant for analysis.

The overall conclusion is that the level of incompleteness and deficiency in the BBD Database is low and fairly acceptable. The database is a remarkable achievement of all those who have worked on this project. Further activities related to this source should be concentrated on improving the quality of information and on enlarging its size by checking the records not yet marked as complete, which task although time-consuming is certainly feasible. A validation of the BBD Database with other sources such as for example, the 1991 Population Census, reporting on the Bosnian population at the outbreak of the 1992-95 war, or the FIS and RS Mortality Databases 1992-95, would be most desired.

PART II : PRESERVATION OF ORIGINAL INFORMATION

2.1 INTRODUCTION

Together with the completeness of data found in the database, the team also verified the source material used in creating of the database and the way information was translated from the sources into the database. Given the enormity of the source material, this job could only be undertaken on a sample basis.

2.2 METHOD USED

From the database of approximately 96.000 active cases, a random set of 50 cases (0.05%) was drawn. For these cases, the entire data entry process was thoroughly checked. Leading questions were: what sources were used? What was the coding protocol? What decisions were made in the event of conflicting evidence between sources? What information was not entered or lost during the coding. In the evaluation, we looked into the sources used in the data entry process and checked all entries.

The evaluation started by sorting the entire database of active victims alphabetically. In the version the consultancy had received in July 2006 there were a bit more than 96.000 active cases of victims. We used a technique called *systematic sampling*, whereby every 1920th case (96.000/50) was sampled. Given that there is no periodicity or pattern in the database when it is ordered alphabetically, there will be no bias resulting from this technique. Each case in the database has a known an equal probability of selection. In order to avoid any bias, we did not sample the first, the 1920th and so on cases, but drew a random number between 0 and 1920 (=96.000/50) and added that number 50 times with 1920. This looks as follows

(1) step one: draw X randomly between 0 and 1920
(2) step two: select X ; X + 1920 ; X + (2*1920) ; ….. X + (49*1920) as case numbers

All the cases that matched the selected case numbers in the file given to the consultancy team were chosen to be in the sample.

2.3 RESULTS

The data base developed by the Centre is a multi-purpose, relational data base. Its use to compute the number of victims is only one of the possible uses. The set-up of the data base allows one to relate data on victims with documents, data on perpetrators, court proceedings and so on. This relational element is very important in the future use of the data base.

For all of the 50 cases, we verified 21 variables. The variables are the following:

- The name of the victim; the name of the father;
- The date of birth; place of birth; code for the municipality of birth;
- The code for nation; code for religion; code for sex;

- The occupation; marital status; name of army (if soldier);
- The date of suffering; code for the kind of suffering;
- Was person found in mass grave?; Status of victim as civilian or soldier;
- The location of the grave; the code for municipality of residence;
- The code for municipality of suffering; citizenship;
- Is the person dead?; the identity of perpetrator

There are several other variables in the data base like the level of education of the victim. However, the education variable was generally not coded, disqualifying it as a candidate for evaluation.

Table 34 presents the results of the evaluation of the 50 cases. From the maximum number of 1050 entries (50*21), 864 have actually been filled.[17] The discrepancy between the maximum number and the actual number of entries is caused by missing data in the original documents. The variables on the data file, 21 for each case, are designed by the data base developers. It cannot be expected that the entries for all of these variables, for all of the cases, would be present in the documents at hand.

From the 186 missing entries (1050-864) we found 5 that could have been entered with information from the documents. Inclusion of these 5 entries would bring the total number at 869 (864+5). The 5 items that were not registered in the database are the information that the victim was married (2), the occupation of the victim (2) and the identity of the perpetrator (1). From these 869, we found 855 that were entered correctly in the database. 5 not entered and 9 mistakes. The percentage of wrong coding is thus 1.61% (1-855/869).

We observe that on average 17.28 variables are entered (864/50) and that on average 17.10 variables are entered correctly (855/50). We can calculate a 95% confidence interval to see whether or not the observed values remain within an acceptable range. Since the mean of the variables is 17.28 (X bar in the formula) and the standard deviation 2.021 (s) the formula for a two-sided 95% confidence interval is

$$P\left(-1.96 \le \frac{\overline{X}-u}{\frac{\sigma}{\sqrt{n}}} \le +1.96\right) = P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le u \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) =$$

$$P\left(17.28 - 1.96\frac{2.021}{7.07} \le u \le 17.28 + 1.96\frac{2.021}{7.07}\right) =$$

$$P(17.28 - 0.56 \le u \le 17.28 + 0.56) = P(16.72 \le u \le 17.84)$$

[17] For civilians, the maximum number of entries is actually 20, not 21 since the army variable is missing by definition. 864 is the total number of non-missing entries for the 21 variables mentioned above.

Which means that if we would draw a similar sample of 50 cases numerous times from the database and count the number of average variables that were entered, then in 95% of the cases we would find a result between 16.72 and 17.84.[18] The values found for the average number of variables that should have been entered (869/50=17.38) and the average number correctly entered variables (855/50=17.1) are all within the 95% confidence interval. [19] In fact, we were able to calculate the average number of variable entries (non-missings) per case for these 21 variables in the entire database. The result, 1,673,620 entries for 96,895 active cases is 17.27 on average, which is almost exactly the same result as in our sample of 50 cases.

Among the 855 entries we found 9 mistakes (864-855). One entire case was a duplicate. This is a mistake since a case should be registered only once as an active case in the database. 3 entries mentioned 'murder' as the cause of death, whereas the documents at hand mention 'killed by sniper, explosives (or shelling) and slaughtered' respectively. These 3 victims are of course also murdered, but the database has separate entries for the cause of death, including killed by a sniper, an explosive, or slaughtered. Since the database allows such fine-grained analyses on the cause of killing, we count these entries as mistakes. 1 person was coded as missing, but he is actually dead. This miscoding can be explained by new information from exhumations or other sources that has not been added to the database. 2 cases had the wrong code for the municipality where the victim died, for one of them also the name of the municipality was not correct. The 2 remaining mistakes were the wrong status of a victim (soldier in stead of civilian) and the date of death being registered in the database but intractable in the documents.

During the evaluation considerable difficulty was experienced with the format of dates in the data base, being a variable of the form 10-8-1970 for a person born on the tenth of August 1970. This caused difficulty because the day, month and year were all part of one and the same variable 'date of birth' or 'date of suffering' . When, example given, only the year of birth is mentioned in the documents (1965), the database registers this person as 1-1-1965 and the user does not know whether the day and month are missing or whether the person is actually born on the first of January. Alternatively, when day or month are missing, it is possible that the date, including the year is not registered at all. This problem can easily be solved by having three separate variables for day, month and year.

2.4 CONCLUDING REMARKS

This evaluation found that 855 of the 869 variables (or 98.39% of the possible entries) were entered correctly in the data base. This corresponds to on average 17.1 variables per case, which falls within the 95% confidence interval. This means that if we would draw a similar

[18] In the formula of the confidence interval, we used s the sample standard deviation in stead of the population standard deviation (sigma). Since the latter is unknown, we use s, which is close to sigma if n is large enough. In general s is an underestimation of sigma.

[19] When calculating a one-sided 95% confidence interval to see whether or not the 17.1 correct entries fall within a reasonable distance of 17.28, the value of 16.81 is calculated as the lower bound of the interval, meaning that 17.1 is found to be in the interval. This is also the case when using 17.38 as the entered average. In that case the lower bound is 16.91 meaning that 17.1 is still within the 95% interval.

random sample of 50 cases many times from the database and count the number of average variables that were entered and that were entered correctly, then in 95% of the times we would find a result between 16.72 and 17.84. Meaning that the entry of variables in the database occurred with very few mistakes.

The mistakes found include a duplicate, which is the most serious mistake. The other mistakes include lesser mistakes such as the wrong municipality of the victim or a different code for the cause of death. Importantly, the team in the Centre is very aware of the remaining inaccuracies (such as the entry of the dates) and is constantly improving and updating the data base.

Table 34. Checking the Entry of 50 Cases in the Database

A	B	C	D	E	F
Nr of case	**Nr of documents used**	**Nr of Fields filled (max 21)**	**Which empty field(s) could have been entered with data from the documents**	**Nr of Fields entered correctly (max is in C)**	**Which mistake made in fields that were entered?**
1	4	19	Occupation is unemployed	18	Code of suffering 9 should be 11
2	1	15	-	15	-
3	2	18	-	18	-
4	1	13	-	12	Duplicate, other active case spelled with one letter difference and linked to 4 other documents
5	1	17	-	17	-
6	2	16	Occupation is retired	16	-
7	3	18	Marital status is married	17	Code of suffering 9 should be 6
8	1	11	-	11	Date of birth not registered because day and month of birth missing in the documents
9	3	15	-	15	-
10	4	20	-	20	-
11	1	18	-	18	-
12	2	17	-	15	Status should be civilian,

					Municipality of suffering should be 12
13	2	20	-	20	-
14	2	17	-	17	-
15	3	20	-	20	-
16	1	17	-	17	-
17	2	15	-	15	-
18	2	14	-	14	-
19	2	17	-	17	-
20	4	19	-	19	-
21	2	16	-	16	-
22	2	18	-	18	-
23	3	16	-	16	-
24	2	16	-	15	Municipality of suffering should be Bihac
25	4	19	-	19	-
26	3	19	-	19	-
27	1	16	-	16	-
28	2	19	-	19	-
29	8	18	He was married and the identity of the perpetrator is also documented: Ustasha	17	Code of suffering was 2 (slaughtered), not 9 (killed)
30	4	16	-	16	-
31	3	17	-	16	Person is not anymore missing but he is dead (code of suffering is 9)
32	6	16	-	16	-
33	3	17	-	17	-
34	2	17	-	17	-
35	1	18	-	18	-
36	6	20	-	20	-
37	1	16	-	16	-
38	1	17	-	17	There are several other doc. related to persons from the same village with the same name, thereby easy to confuse
39	1	14	-	14	-

40	6	18	-	18	-
41	2	16	-	16	-
42	10	20		19	Date of death was registered in database, but could not be found in the documents
43	3	21	-	21	-
44	2	17	-	17	-
45	3	18	-	18	-
46	2	20	-	20	-
47	1	19	-	19	-
48	2	16	-	16	-
49	2	20	-	20	-
50	1	18	-	18	-

PART III : ANALYSIS OF DEDUPLICATION AND COVERAGE[20]

3.1 INTRODUCTION

Finding duplicate records (in one or several databases) is very hard both for people and for computers.[21] Names may be reported in slightly different spellings or abbreviations in different records, dates may vary, and places may be described or coded in slightly different ways. Furthermore, data may be missing from some fields.

In the other direction, a database of events in the world is necessarily partial. Although there are an unprecedented number of deaths documented in the BBD, there are certainly other deaths which have not been documented by the BBD. For example, there may be many deaths in other databases which are not included in the BBD, and there may be some deaths which have never been documented in any database.

All of these problems affect the BBD records: the BBD contains undetected duplicates, and there are deaths which are not documented in the BBD. In Part I of this report, pairs of records among the active set were identified that matched exactly in combinations of fields and partial fields, finding a minimum of approximately 1000 undetected (latent) duplicate pairs. In Part II, one of the 50 sampled records was found to be an undetected duplicate, implying that there are approximately 2000 latent duplicates.[22]

All record linkage techniques depend on a set of rules that define whether records are duplicates or not. In manual techniques, the database staff must evaluate all potential duplicates and make decisions either by explicit rules or by intuition. There are many problems with manual deduplication, the most severe of which are:

a) lack of transparency: it is impossible to examine the database and determine either what the deduplication rules were;
b) lack of consistency: even if the rules were explicit, it is impossible to tell if they were applied consistently by different staff members and by the same staff member at different times;
c) lack of replicability: if the organization wants to change the matching rules, there is no way to re-apply the rules except by re-doing the entire project.

This section explores the duplication problem using automated record linkage techniques which are designed to mitigate these three problems. Computer-based deduplication techniques examine small "training sets" of candidate pairs that have been labelled by people who understand the database ("experts") as either matches or non-matches. Using a wide array of comparisons, the computer models deduce the rules applied by the experts. Through

[20] Jeff Klingner of Benetech's Human Rights Data Analysis Group adapted external libraries, wrote the software used in the deduplication analysis, conducted the deduplication, and contributed to the reasoning in this section.
[21] The classic work in this field is by I.P. Fellegi and A.. B. Sunter, "A Theory for Record Linkage," *J Am Stat Ass*, 64(328):1183-1210. 1969. See also W. Winkler, "The state of record linkage and current research problems," at http://citeseer.ist.psu.edu/article/winkler99state.html as of 14 June 2007, and T.N. Herzog, F.J. Scheuren, and W.E. Winkler, *Data Quality and Record Linkage Techniques*, Springer-Verlag: 2007.
[22] (1/50)*96895 = 1938, in a confidence interval [0-5736].

additional iterations, the expert adds additional training sets and re-calculates the model until the model can classify the training pairs in the same way as done by the expert.

It is important to understand that computer-based deduplication is no more than the consistent application of matching rules deduced from examples classified by people. Ultimately, the rules come from people, not from the computer. Indeed, automated techniques make it possible to compare the impact of using slightly different sets of rules or deduplication criteria (such as those from different experts) to assess the sensitivity of the data to the construction of the training set. Throughout this Part, we will suggest different mechanisms by which duplicates in the BBD could be detected, and we will speculate about how those rules would likely affect the results.

3.2 DEDUPLICATION AND SELECTION BIAS

Among this report's recommendations is that the BBD can be used for basic descriptive statistics. Descriptive statistics usually means to analyze and interpret the database (e.g., "deaths known to the RDC and documented in the BBD"), in contrast to trying to understand patterns in the real world (e.g., "deaths that occurred during the conflict"). The BBD can be used to make descriptive statements of the former kind, such as "there were, at minimum, approximately 96,000 deaths in the conflict." Similarly, the statement that 90% of the deaths documented by the BBD were males is accurate. However, without further statistical treatment, the BBD could not be used to make statistical statements about general patterns in the conflict.

There are significant *potential* problems with using the BBD – or any database – to make statements about the real world.[23] Our interest in data is usually to find patterns, by which we mean to find similarities or contrasts between subcategories in the data. For example, imagine that we want to understand the rise and fall of deaths over time, measured by month. We want to understand this pattern about the real world, that is, we want to generalize the pattern found in the database to the universe. However, to make this generalization, the comparison of each month with the others must assume that no month is especially affected by either latent duplication or undocumented deaths.

To illustrate the problem, imagine that 1% of the records in the BBD are duplicates, and that all the duplicates have a date of suffering in March 1993. A comparison of March 1993 with other months would be biased because March 1993 would have double-counted some 1000 deaths, whereas other months did not. Imagine further that there are only 1000 deaths that occurred during the conflict but that have not been documented by the BBD. If all of these deaths occurred in April 1993, the trend between March and April could be completely determined by data limitations (latent duplicates and undocumented deaths). This problem

[23] Here we use "database" to mean a list that was not collected by a probabilistic ("random") method. Databases are large convenience samples, usually partial or attempted censuses. In this case, we mean that the RDC did not draw a random sample: they intended to enumerate *all* of the deaths in the conflict, even as they recognized the impossibility of reaching that goal.

might occur if the duplicates or undocumented deaths are concentrated in any dimension of interest in the database (e.g., space, ethnicity, type of event, etc).

To make a valid statistical inference about the universe, comparing results across subcategories in the BBD, we must solve both the deduplication problem and the selection bias problem. The representation problems (latent duplicates and undocumented deaths) are interdependent because the statistical technique used with databases to make projections and inferences relies on information about the density of linkages among duplicate records. That is, perfect matching is a central assumption of the estimation technique called multiple systems estimation. [24]

If we can assume or estimate a list of possible duplicates, we can assess the potential bias caused by latent duplicates. As with selection bias, it is not possible to evaluate the impact of the latent duplicates without solving the problem, that is, identifying the duplicates. In section 3.6, below, we show that the bias due to likely latent duplicates does *not* substantially affect the results of descriptive statistics for the BBD. The inferences, however, would be strongly affected.

The remainder of this Part considers these problems empirically, by evaluating the existing matching, estimating the latent duplicates, evaluating the sensitivity of descriptive analysis to the latent duplicates, and recommending steps to complete and improve the BBD.

3.3 ASSESSING EXISTING MATCHING

We assessed the existing matching by estimating a decision tree. The model was created to distinguish random pairs of records selected from the active and inactive sets from pairs of records (one active, one inactive) identified by RDC as duplicates.[25,26]

---

[24] For the mathematical foundation of multiple systems estimation; see Y.M. Bishop, S. E. Fienberg, and P. H. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. 1975, especially chapter 6. Also see S. E. Fienberg, M.A. Johnson, and B.J. Junker, "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society*, Series A, 162 , 383-406, 1999. On model selection, see J. A. Hoeting, D. Matigan, A.E. Raftery, and C.T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science* 14(4):382-417, 1999. For an application similar to the BBD, see P. Ball, W. Betts, F. Scheuren, J. Dudukovich, and J. Asher, "Killings and Refugee Flow in Kosovo March – June 1999: A report to the International Criminal Tribunal for the Former Yugoslavia," Washington DC: AAAS/ABA. 3 January 2002 at http://shr.aaas.org/kosovo/icty_report.pdf as of 14 June 14, 2007.

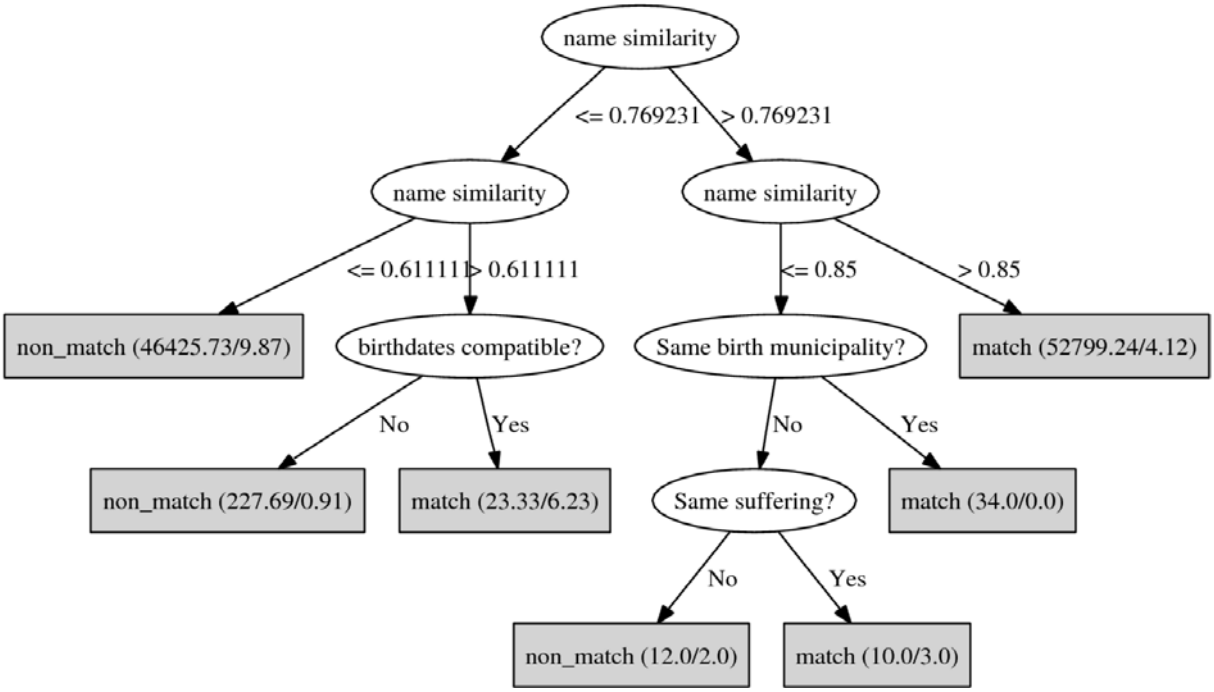
[25] The data used were the July, 2006 data and matching information provided by the RDC 27 April 2007.

[26] For a discussion of the decision tree model and other techniques described in this section, see Ian Witte and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (second edition), Morgan Kaufmann (now Elsevier), 2005. A summary of the challenges in database deduplication and clustering can be found in Mikhail Bilenko, "Learnable Similarity Functions and Their Application to Record Linkage and Clustering," University of Texas at Austin, Department of Computer Science Technical Report UT-AI-TR-03-305, 2003. Other software implementation techniques are described at the website of the Australian National University Data Mining Group, http://datamining.anu.edu.au/projects/linkage-publications.html, as of 14 June 2007.

Among the inactive records, there were 51,870 records linked to active records.[27] In addition, there were 27,428 pairs transitively implied by those positive hand matches.[28] We added 70,000 randomly selected pairs as likely negative matches so that roughly equal numbers of positive and negative matches were available for the model estimation.[29] In total, these positive and random examples are called the candidate pairs.

The decision tree below fits the hand matching in the candidate pairs very well: it can correctly distinguish hand matches from random pairs in 99.98% of the cases.

Figure 7. Decision Tree to Match the July 2006 Data



To interpret this model, imagine a pair of records. The model will classify this pair of records as either matching or not matching by following the rules defined in the tree, starting at the root of the tree (the single node at the top of the diagram). The circles indicate decisions: if the names were similar (i.e., differing by only a common misspelling, shown by the weight >0.85), then the record pair would travel the rightmost branch to the "match" outcome. The reported statistics (52799.24/4.12) indicate that 52799.24 candidate pairs that were indicated as matches were correctly classified as matches by the model (fractional pairs will be discussed below). In contrast, 4.12 records have very similar names, yet were labelled as non-matches by the original coding.

The number of records found in the match and non-match outcomes can be fractional. When pairs of records are evaluated (such as by measuring whether two records have the same birth municipality), records that are missing data for the field being evaluated are divided, with

[27] Matched records are those for which id & active_id are not equal, and are both valid ids.
[28] For example, if records A and B are inactive, and they both are linked to active record C, transitivity implies that A and B could also be linked.
[29] Transitively-closed hand matches and candidate pairs were excluded from the selection.

fractional records passed down both branches of the decision point. This allows incomplete records to be included in the matching analysis.

The model is presented to introduce the specific deduplication recommendations. However, there are three problems with using this model directly to determine the extent of over- or under-matching:
1. Names are perfectly similar in hand matches. We suspect that after matches were identified, RDC staff often edited the duplicate records to be the same ("correcting" the active records). The editing changes the underlying data, so our model was built not on the original data the RDC staff examined. Instead, the model was built on corrected data. As a result, this model underestimates latent duplicates because it cannot detect them among records with substantial differences in field values.
2. Not all the linkage information from the inactive records is available: very few inactive records have their active counterpart listed. This lack of information means that the model was built only on a fraction (approximately 1/3) of the hand matching. We do not know the effect of this problem on estimation of latent duplicates by the decision tree model.
3. The training set only has positive training pairs. Ideally, the model should have specifically labelled non-match pairs to compare against the pairs labelled as matches. Given the resources available for this analysis, we could not create negative examples.

3.4 EVALUATION OF THE HAND MATCHING: JULY 2006 DATA

As reported above, both Parts I and II of this report found or inferred that the BBD contains latent duplicates. This section considers the same question from a slightly different perspective. This section uses the July 2006 data, which is the same data as used in the analyses in Parts I and II; section 3.5 applies many of the same tests to data current as of May 2007.

We followed a method similar to the equality checks described in Part I. We created a comparison of all the active records to each other, including all the fields in the comparison. The comparisons were of three types: equality (are the fields equal?); approximate string comparisons (how many editing operations are required to make the strings equal?); and date overlaps (how far apart are the dates?). The resulting comparison measures were merged in a weighted sum. The comparisons are shown below in Table YY.

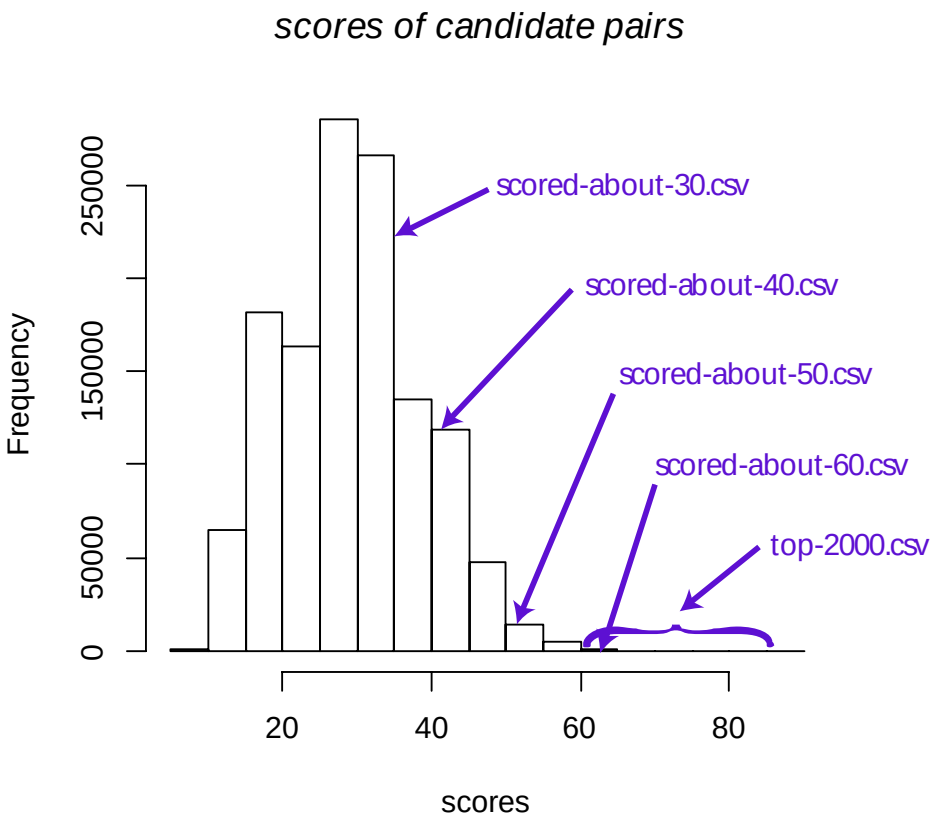
Table 35. Comparisons Used to Rank Record Pairs as Probable Duplicates (In Declining Order of Importance)

Field	**Comparison**
name	Approximate String Comparison
uid	Approximate String Comparison
birth_date	Equality
birth_date	Date overlap
suffering_date	Equality

suffering_date	Date overlap
fathers_name	Approximate String Comparison
Birth_Municipality	Equality
Address	Approximate String Comparison
Religion	Equality
Nation	Equality
Sex	Equality
Occupation	Equality
Suffering	Equality
Citizenship	Equality
Dead	Equality

The histogram of scores in the ranked list of likely duplicate candidate pairs is shown below in Figure ZZ.

Figure 8. Scores of Candidiate Pairs



The scores have an arbitrary scale, and they should be interpreted as no more than an ordered ranking of the similarity of the pairs, combined across all fields. Pairs with high scores are more similar than pairs of lower scores. In our subjective assessment, most (if not all) of the 2000 candidate pairs with the highest scores are true duplicates.[30] However, true pairs continue to appear with scores lower than 60, though they become less frequent with

[30] This is consistent with the estimate of approximately 1000 duplicates made in Part 1

declining scores. There is no specific threshold which divides matches from non-matches in the scoring, so we cannot estimate the true number of latent duplicates in the active set.

3.5 IMPLICATIONS OF THE HAND MATCHING DECISIONS: MAY 2007 DATA

By May 2007, the RDC had produced a new version of the BBD with 97,207 active records and 34,378 inactive but linked records. We repeated the process described in section 3.4, above, with this new data and with more flexible comparison rules.

Table 36. Comparisons Used to Rank Record Pairs as Probable Duplicates

Variable	**Comparison**
Name	Approximate String Comparison
Uid	Approximate String Comparison
Date_of_birth	Frequency based match weights
birth_date	Date overlap
Date_of_death	frequency based match weights
death_date	Date overlap
Fathers_name	Approximate String Comparison
Municip_birth	equality
Address	Approximate String Comparison
Religion	Equality
Nationality	Equality
Sex	Equality
Municip_of_permanent_address	Equality
Municipality_of_death	Equality
Citizenship	Equality

In this analysis, we introduced frequency-based match weights. This form of comparison treats matches of field values that are common in the data as providing less information than matches of field values that are rare. In the context of the BBD, a pair of records that both have a date of death during mid-July 1995 (when there were many deaths, especially due to Srebrenica) tells us less about whether they are the same than if the pair of records both had a date of death in mid-July 1994 (when deaths were relatively less frequent).

With the analysis of likely pairs in this analysis, we excluded the "equality matches" described and evaluated in Part I of this report. The equality blocking techniques find pairs of records which are equal on a set of fields. In order to find possible matches with non-identical names, the equality check is often performed on name prefixes (the first letter or the first three letters, called "stemming"). These techniques are useful, but they miss duplicate records with names that differ in their first few characters.

One example of such differences is the common alternate spelling of DŽ/Đ (Latin D followed by Z-with-caron vs. D-with-stroke), which often occurs at the beginning of a name (e.g.

DŽONLIĆ vs ĐONLIĆ). There is a similar issue for Ć/Č, though this variation is less common. String comparisons based on common endings can be handled by reverse-stemming, but the variations in string comparison tend to be greater than can be captured by strict equality comparisons.[31]

An important difference between the two techniques is that equality blocking requires matches on many fields to find possible duplicates, while the similarity model is able to rank possible matches based on many field comparisons. This lets the similarity model find duplicates that differ only slightly or have missing values on multiple identifying fields (e.g., name, uid, birthdate, death date).

Among the top 5000 scoring pairs of potential duplicates, excluding pairs that would have satisfied the equality criteria described in Part I, we find several thousand possible additional duplicates. Without expert guidance from the RDC, it is impossible to be more precise about how many of these are true duplicates; in section 3.7, below, we recommend that the RDC process these records further, along with the equality matches identified in Part I, before conducting descriptive analysis.

3.6 SENSITIVITY ANALYSIS OF DESCRIPTIVE STATISTICS TO LATENT DUPLICATES

Descriptive statistics from the BBD are not substantially affected by bias due to undetected duplicates. We considered five key variables (nationality, religion, sex, municipality of death, and date of death). For each variable, we calculated the tabulation of values for the complete set of active records (exactly like the tabulations presented in Part I). We calculated a second tabulation based on the active records with the 8000 highest-scoring (most similar) candidate pairs deleted. We found that the largest shift among categories with up to 8000 potential duplicates removed was for all variables less than 2%.

Consequently, we conclude that descriptive statistics of the BBD are unbiased with respect to likely latent duplicates. Inferential statistics, however, tend to be quite the reverse: multiple systems estimation assumes that deduplication has been perfect (or nearly so). The estimate is derived from the pattern of duplication among databases, and consequently, the estimate can be strongly affected if the identification of the pattern (i.e., the deduplication) is not accurate.

3.7 RECOMMENDATIONS FOR CONTINUING CORRECTIONS TO LATENT DUPLICATION AND UNDER-REGISTRATION

The original intent of this section was to make a multiple systems estimate of the total number of deaths not documented in the BBD. This technique uses the density of duplication across multiple lists to make a statistical inference about the number of elements not on any of the

[31] For a review of string-difference measurement techniques, see M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S.E. Fienberg, "Adaptive Name Matching in Information Integration," *IEEE Intelligent Systems*, 18(5): 16-23. 2003.

lists in the data. The nature of the BBD as a combination of multiple individual sources is the appropriate starting point for multiple systems analysis. However, the unavailability of the linkage information among the inactive records makes it currently impossible to estimate the total.

Our goal is to recommend techniques that will enable the RDC to identify the records that compose the best representation of unique deaths during the war. With improved deduplication, it will be possible to estimate the total deaths due to the conflict.

In increasing order of complexity, effort, and benefit, we recommend the following:

1) Complete deduplication among records in the active set using flexible comparisons. Although the descriptive analysis is not substantially sensitive to the likely latent duplicates, we nevertheless recommend that the deduplication be completed before descriptive analysis. This deduplication could use a combination of the equality tests in Part I and the ordered pairs presented in section 3.5 to continue eliminating duplicates in the current active set.

2) Reconstruct the linking between inactive duplicates and the active set. If there is interest in correcting for under-registration (i.e., for estimating statistically the number of deaths not recorded in the BBD), the first step is to determine the complete link information between inactive and active records. In the latest linkage information sent by RDC, records have the following status:
97,207 active records
34,378 inactive, linked records (duplicates identified with the active record they link to)
118,520 inactive, unlinked records
250,105 total records

The 118,520 inactive, unlinked records are currently excluded from analysis due to a combination of a) missing link information that would indicate that some of these records are duplicates of active records; and b) highly incomplete event information (especially date of death) preventing the RDC from confirming the reports. In order to conduct a reliable inference using multiple systems estimation, all the records that can be matched, should be matched. If the original link information is available in earlier versions of the database, this information could simply be recovered. Alternatively, computer based matching techniques could be used to produce and process training sets that would supplement the existing hand matching. Once the linking is reconstructed, multiple systems estimates could be made. This alternative assumes that the existing hand matching (the 34,378 inactive, linked records) is accurate. We caution that the deduplication that can be done based on the existing hand matching in the 34,378 inactive, linked records is risky for the reasons explained at the end of section 3.3.

3) Use the unedited records to construct a new deduplication model. The most thorough approach would be to recover the original unedited field values. If there are backup copies of the database available from previous years, the field values in active and inactive records could be restored to their pre-edited (uncorrected) state. A new deduplication model could be constructed using the full variation of the unedited records, and including the records with

missing information. This approach is the most likely to estimate deduplication accurately, and thereby serve as the most rigorous basis for inferences about the total deaths during the conflict.

4. OVERALL CONCLUSION

Any database on war victims contains deficiencies like data entry errors, duplicates, missing values; its coverage is incomplete but usually also partly overlaps with other databases. All these problems were anticipated for the Bosnian Book of Dead Database, a unique source that has been established and maintained since many years by the Documentation and Research Centre in Sarajevo.[32] What distinguishes the BBD database from other sources is its size; it contains about 96,895 records in just one "active" data table and another 149,841 records that have been used for checking, corrections and possibly inclusion into this "active" table. This size offers a tempting opportunity to estimate numbers of victims of the Bosnian war based on just one single source. Whether or not one may use this source for this kind of estimation is not only a matter of the balance between the size and deficiencies of the database, however. Even though this balance is a key for using the database at all, reconstructing the history of human suffering in the Bosnian conflict in 1992-95 on the basis of one single source might result in erroneous conclusions. Next to data deficiencies such as data entry and other errors, duplicates and missing values, there are biases related to methods of data collection and quality of reporting inherent in this data too. These biases cannot be easily detected by studying the database itself; comparisons with other sources in their original form are invaluable in assessment of this kind of biases. The possibilities for comparisons, although limited, do exist and we encourage the authors of BBD to engage in comparative studies.

In this context it is useful to note that sources on victims of the Bosnian war are generally extensive and include, for example:

- The FBH 1992-95 Mortality Database established in 2002 by the Federal Institute for Statistics in Sarajevo. (About 25,000 war-related and 50,000 natural death records).
- The RS 1992-95 Mortality Database compiled in 2005 by the Statistical Office of Republika Srpska in Banja Luka. (About 16,000 war-related and 50,000 natural death records).
- The ICRC list of missing persons. (About 22,000 records).
- Several other lists of missing persons including those by the FBH and RS Commissions for Tracing Missing Persons, another one by the International Commission for Missing Persons (ICMP) in Sarajevo, and several lists published locally (like for Prijedor and other municipalities).
- Official military lists of fallen soldiers and military and police personnel of the FBH and RS Ministries of Defence. (About 50,000 records)
- Records of the exhumed and identified persons in possession of the FBH and RS Commissions for Tracing Missing Persons, and of the ICMP. (The persons identified through the DNA matching methodology alone amounted recently to at least 8,000 individuals in Bosnia).
- Sarajevo Household Survey of mid-1994. (About 6,000 war-related deaths in Sarajevo until mid-1994).
- Many other lists by various NGOs.
- And of course, there is the Bosnian Book of Dead Database.

[32] Formerly, the BH State Commission for Gathering Facts on War Crimes, chaired by Mirsad Tokača.

Each of the above sources is indispensable in answering specific questions meant to be answered by this given source. However, when it comes to statistics on victimization of a war, none of the above sources, if used alone, can be seen as sufficient. None of them can be then considered complete and fully unbiased with respect to statistics on victims of the 1992-95 war in Bosnia. The BBD is by far the largest, most complete and most complex source in this context, and therefore the most encouraging to use. But the best approximations of the truth will be always obtained from results coming from many different sources and many different methodological approaches.

Having studied the 2006 version of the BBD extensively for the needs of this assessment and realizing a striking improvement of the 2006 version when compared with earlier versions of the database, we are happy to be able to recommend the use of the BBD for the following purposes:

- Advancing the reconciliation process in Bosnia and Herzegovina by displaying transparent and methodologically correct statistics on victims of BiH war.
- Propagating the approach and methodology used for the establishment of BBD. When presenting statistics, stressing the need of distinguishing between the minimum numbers and more complete estimates.
- Propagating comparisons of BBD with other sources on victims and additional sources on incidents and episodes of the war for the purpose of a better insight into the historical truth.
- Using the BBD Database for education of young researchers who can apply this knowledge in their careers.
- Using the database for lead purposes in investigative stages of trial preparation in international and/or national courts for IHL violations.
- Using the BBD database for academic research purposes, including expert analysis and testimonies for judicial proceedings.

The database is a unique and valuable source and deserves a prominent place among sources on victimization of the 1992-95 war in Bosnia and Herzegovina.

List of Abbreviations:

Organizations:
ICTY – International Criminal Tribunal for the former Yugoslavia
OTP – Office of the Prosecutor, ICTY
RDC – Research and Documentation Centre, Sarajevo

Database items:
BBD - Bosnian Book of Dead, alternatively known as the Population Loss Project
JMBG - personal identification number
(cl) – marker of cleaned items (e.g. YoB(cl))

Related to Birth:
DoB- date of birth
DayB- day of birth
MonthB- month of birth
YoB- year of birth; the same as YearB
YearB- year of birth; the same as YoB
PoB- place of birth
MoB- municipality of birth

Related to Death
DoD- date of death
DayD- day of death
MonthD- month of death
YoD- year of death; the same as YearD
YearD- year of death; the same as YoD
MoD- municipality of death
CoD- cause of death

Related to Residence:
MoR- municipality of residence

Related to Names:
FstName- first name
SurName- surname or family name; the same as FamName
FamName- surname or family name; the same as SurName
FaName- father's name
(1)- initial of a name (e.g. FstName(1))
(3)- three first letters of a name (e.g. FstName(3))

PROFILES OF THE AUTHORS

PATRICK BALL

Patrick Ball, Ph.D., is the Chief Technical Officer of the Benetech Initiative. He also directs Benetech's Human Rights Program, which includes the Martus project and the Human Rights Data Analysis Group (HRDAG). He received his Ph.D. in sociology from the University of Michigan in 1998. Since 1991, Dr. Ball has designed information management systems and conducted statistical analysis for large-scale human rights data projects used by truth commissions, non-governmental organizations, tribunals and United Nations missions in El Salvador, Ethiopia, Guatemala, Haiti, South Africa, Kosovo, Sierra Leone, Sri Lanka, Perú, Timor-Leste, Bosnia and Herzegovina, and Chad.

Dr. Ball has received several awards for his work. In April 2005, the Electronic Frontier Foundation presented him with their Pioneer Award. Dr. Ball received the Eugene Lawler Award for Humanitarian Contributions within Computer Science from the Association for Computing Machinery (ACM) in June 2004. In August 2002, the Social Statistics Section of the American Statistical Association presented Dr. Ball with a Special Achievement Award for his work on Kosovo.

Dr. Ball is currently involved in Benetech projects in Sri Lanka, Colombia, Burma, Liberia, Guatemala and in other countries around the world.

EWA TABEAU

Ewa Tabeau graduated in econometrics and statistics (1981) and obtained her Ph.D. in mathematical demography (1991) at the Warsaw School of Economics (WSE) in Poland. Several years she taught statistics and demography to students at WSE. She was as well a researcher at the Dutch Interdisciplinary Demographic Institute in The Hague (NIDI), with modeling and forecasting of mortality by cause of death and analysis of prospects for life expectancy being her main research domains. She was then also invited, as an expert, by organizations such as Eurostat – Statistical Office of the European Union; ING Group - Life Insurance NL, Goldman & Sachs - Life Insurance USA, Statistics Netherlands, British Government Actuary's Department, to consult their projects involving issues of mortality and health development and prediction.

Since 2000, Ewa Tabeau has been the project leader of the Demographic Unit at the Office of the Prosecutor OTP), International Criminal Tribunal for the former Yugoslavia (ICTY), in The Hague, where she has studied demographic consequences of the conflicts in the former Yugoslavia and provided crime statistics to trials and investigations at the OTP. The main subjects of her research included statistics and estimates on war-related deaths (killed, missing, exhumed, identified etc.), wounded persons, and on internally displaced persons and refugees. During her employment at the OTP, she completed more than 20 expert reports including, among others, the reports prepared for the Slobodan Milosevic (Bosnia), Biliana Plavsic

(Bosnia), Momcilo Krajisnik (Bosnia), generals Galic and D. Milosevic (the siege of Sarajevo), Blagojevic and Popovic et al. (Srebrenica), and Prlic et al. (Herzeg-Bosnia) cases, and she testified many times as an expert witness before the Tribunal.

Ewa Tabeau authored 5 monographs published internationally, 25 articles published in international and national journals, 19 conference papers presented at international conferences, and more than 60 research reports and working papers.

She supervised researchers completing their M.Sc. and Ph.D. dissertations, and acted as peer reviewer for scientific journals and publishers, such as for example the *European Journal of Population, Journal of Peace Research, Mathematical Population Studies*, Springer, Thela Thesis etc.

PHILIP VERWIMP

Philip Verwimp studied economics and sociology at the universities of Antwerp, Leuven and Göttingen and undertook graduate work in political economy and genocide studies at Yale University. He obtained his PhD in Economics (2003) from the Catholic University of Leuven (Belgium) with a dissertation on the political economy of development and genocide in Rwanda.

Philip taught research methods and development economics at the universities of Leuven, Antwerp, Utrecht and at the Institute of Social Studies in The Hague. He worked as a Poverty Economist for the World Bank (2004-2005) and was a Fulbright-Hays Scholar at Yale University (2004) and a recipient of a Fellowship from the Belgian American Educational Foundation for doctoral research at Yale in 1998-1999. Other scholarships include a 4-year doctoral scholarship from the Fund for Scientific Research and a grant from the King Baudoin Foundation.

In 2004 he was awarded the Jacques Rosenberg Price of the Auschwitz Foundation for his dissertation and in 2006 he received the award for the best paper published in the *Journal of Peace Research*. His work is published in the *Journal of Development Economics, Journal of Peace Research, Population Studies, European Journal of Population, Journal of Conflict Resolution, European Journal of Political Economy*, among others. Research interests include the political economy of development and conflict, poverty, inequality, demography, dictatorship, human rights, genocide. He regularly reviews papers for academic journals, attends and organises conferences and workshops.

Currently, Philip is the co-founder and co-director of the Households in Conflict Network (www.hicn.org) and the deputy director of Microcon, a large EU research project on conflict.