Capture-Recapture Estimation of War Deaths: Foundations, Assumptions, and Challenges

Nicholas P. Jewell
Division of Biostatistics and Department of Statistics
University of California, Berkeley

Michael Spagat
Royal Holloway College
University of London

&

Britta L. Jewell
Green-Templeton College
Oxford University

Section 1: Introduction Capture-recapture estimation has been used for counting elusive wildlife and human populations since the late nineteenth century. A classical application is to capture fish from a pond, count and tag the fish and then return them to the pond. Later there is a second capture (the 'recapture'). Again the fish are counted but, in addition, a separate tally is made of the number of tagged fish appearing in the recapture. We can then estimate the total number of fish in the pond based on the size of the two catches and the overlap between the two.[1] Intuitively, a large overlap suggests that there are probably few fish in the pond that eluded capture twice whereas a small overlap suggests that probably most fish were never caught. We can use data from more than two captures to further improve the estimate of the total number of fish in the pond.

Although these methods were originally designed to count wildlife populations, their use has been expanded to count elusive human communities such as the number of crack cocaine users in London (Hope et al., 2005), the autistic population in Lothian, Scotland (Harrison et al., 2006), the lesbian population in Allegheny County, Pennsylvania (Aaron et

[1] The standard estimate is the product of the two catch sizes divided by the number of twice-captured fish.

al., 2003), and the World Trade Center Tower population on September 11, 2001 (Murphy et al., 2007; Murphy, 2009) amongst many other applications. Reviews of the method include a special issue of the *Biometrical Journal* (Böhning, 2008), two early introductory papers by the International Working Group for Disease Monitoring and Forecasting (1995a, b), and Chao et al. (2001). Desenclos and Hubert (1994), Hook and Regal (1995, 1997), Papoz et al., (1996), Cormack et al. (2000), and Tilling (2001) all critique the application of capture-recapture methods to human populations.

The specific application of capture-recapture estimation to casualty counting is a more recent development that has now been employed in sufficient examples for the methods to be taken seriously and evaluated. The most prominent applications have been to Guatemala for the period 1960-1996 (Ball, 1999, 2003), Kosovo in 1999 (Ball and Asher, 2002; Ball et al., 2002; Ball, 2003), Peru from 1980-2000 (Ball et al., 2003; Manrique-Vallier and Fienberg, 2008; Manrique-Vallier et al., 2011), Timor-Leste from 1974-1999 (Silva and Ball, 2008), Casanare, Colombia from 1998-2007 (Lum et al., 2010) and Bosnia from 1992-1995 (Brunborg et al., 2006; Zwierzchowsky and Tabeau, 2010). In what follows, we raise some general questions to consider in assessing the application of capture-recapture to the casualty estimation field while not dissecting any particular example in detail. Many of the referenced analyses have, in fact, made efforts to address the issues we discuss below.[2]

There are a variety of nomenclatures for the capture-recapture methodology that are rooted in different contexts and applications. Capture-recapture, or mark-recapture, focuses attention on counting animal populations in the wild. These terms suggest two distinct counting exercises, the capture and the recapture, although both the methods and the terminology have been extended to multiple recaptures. In human populations, estimation is based on overlapping incomplete lists that play the role of the different captures: the techniques are sometimes referred to as dual-record estimation when there are two lists, and multiple systems estimation when more data sources are available. We believe that the classical capture-recapture terminology is the most evocative and intuitive one available, and so we consistently use this term throughout the present paper. Readers should be aware, however, that the literature on casualty

[2] We feel obliged to also point out that none of us have been involved directly in any of the referenced analyses and so are not in a position to comment in detail on any specific case beyond what has been provided in published reports.

estimation generally employs the term ‘multiple systems estimation’ to describe capture-recapture methods.

It is laudable to aspire to produce accurate war-death estimates, bracketed by appropriate confidence intervals to reflect uncertainty. In many conflicts, there is likely to be considerable political pressure to produce such estimates (Andreas and Greenhill, 2010), for example, to focus blame on perpetrators of human rights violations, or to exert pressure on warring groups. However, it may not always be possible to make reasonably accurate capture-recapture estimates, and proceeding in such situations may have the effect of exaggerating the quality of our knowledge. We suggest that in some cases it may be preferable to give credible lower, and perhaps upper, bounds for war deaths based on the lists’ data. These bounds can either supplement, or even replace, estimates based on uncertain and unverified assumptions, and will often be sufficient for practical purposes while directly conveying a sense of necessary imprecision. For example, a war crimes prosecution may benefit greatly from a very well documented lower bound on the war deaths attributable to a particular perpetrator whereas a good estimate of the number of these deaths may arguably add little extra. A poorly founded estimate will only generate confusion.

Another tempting possibility is to try to link variation in counts—over time and geography, for example—to changes in underlying conditions that can be quantified in various ways. Unfortunately, it is sometimes not possible to disaggregate capture-recapture estimates with sufficient reliability to underpin such an analysis. For example, Landman and Gohdes (2011) report that available Peru data are simply too sparse for reliable time disaggregation. In such cases, explanatory analyses using raw counts may still be useful even in cases where these counts are known to be incomplete. If, for example, biases within counts are reasonably well known to be relatively consistent, then at least some useful boundaries on the effects of explanatory factors can be derived. We must, however, sometimes accept that available data will not allow any useful inference.

Many basic questions concerning the application of capture-recapture methods to casualty estimation are simple to state but hard to answer fully. The main questions that we at least broach here include: When, and under what conditions, can capture-recapture estimation provide accurate and useful counts of the number of conflict casualties? Specifically, how many lists are necessary to appropriately allow for list relationships, and how can this question be addressed using available data and knowledge of how the various lists were assembled? The latter involves the issue of whether some lists are best aggregated although such aggregation complicates the understanding of how specific lists were assembled including their interrelationship. In the other direction, there may be a need to disaggregate lists according to specific characteristics including chronological time and geography; that is, to stratify the data appropriately to yield lists that better satisfy the necessary assumptions discussed below. Necessarily, this raises the question of the type and level of stratification that can be most feasibly used. How are the results and assumptions of a particular capture-recapture analysis best translated for political consumption and the media without undermining the credibility of the approach? Ultimately, in each conflict situation, the minimum level of data validity that is required for the methods to be applicable must be clearly answered. Given that the data meet basic quality standards, what methods are appropriate for choosing amongst several capture-recapture estimates with an eye to balancing the transparency of assumptions with statistical constraints? We cannot tackle all of these issues in detail, but we touch on some of them below and challenge researchers to respond fully as these methods continue to be applied to casualty estimation.

Section 2: Basic Statistical Assumptions/Requirements. Capture-recapture estimation of conflict deaths requires assumptions that describe how the lists of deaths (captures) are created, including possible links between these multiple, overlapping and incomplete lists of victims. As in all statistical procedures, the correspondence between reality and assumptions is crucial to the validity of final estimates and their interpretation. In this section, therefore, we examine the assumptions underlying capture-recapture estimation.

We assume that researchers have more than one distinct list (for convenience, let *k* stand for the number of lists), each list providing information and identifiers on casualties in a certain well-defined region for a specified period of time.[3] The basic assumptions underlying capture-recapture estimation are loosely summarized as follows:

[3] With only one list, extremely strong and unverifiable assumptions are required to extrapolate meaningfully from who is on a single list to who is not on the list (Böhning and van der Heijden, 2009).

1. **Coverage** We first assume that the population of victims is closed, meaning here that every death has a positive probability of discovery by at least some list. In the wildlife application this assumption is usually interpreted as meaning no immigration or emigration, in part because it is assumed, for example, that any fish in a pond can be captured unless they exit the pond or have not yet arrived before capture attempts. However, in conflict situations there may be deaths that will never be observed no matter how many lists are compiled. No statistical cleverness can uncover such deaths. We can only estimate the number of deaths that could have been captured but, *by chance*, were not.

Coverage may be most difficult for real-time counts--due to rapid changes in populations in the midst of conflict, and difficulty in immediate casualty ascertainment--but may improve over time for historical estimates. There is, however, a real concern that casualties amongst families that have become refugees may be lost to discovery by many lists. In this regard, it is particularly important to explicitly define the population of victims that is being considered.

2. **Accuracy**

(i) **Perfect matching** It is generally assumed that matching across different casualty lists is perfect. That is, matched deaths truly are the same deaths, and deaths that are not matched truly are different deaths (this includes that no two casualties on the *same* list are duplicates).

(ii) **No false deaths** Deaths appearing on lists are assumed to be real.

3. **Homogeneity** The simplest homogeneity assumption is most easily articulated by requiring that each casualty (from the unobserved total list of casualties) has an equal probability of appearing on a specific list, and this is true for all lists. There is no requirement that these probabilities be equal across lists. So some lists may be more comprehensive (high probability of capturing a casualty), and some much more sporadic (low probability of capturing a casualty). As discussed in further detail below, an implication of homogeneity is that deaths that do not appear on any list do not possess inherent characteristics that make them fundamentally different from deaths that made it onto one or more lists, i.e. it is only due to pure chance that some deaths get discovered and others do not. Essentially, it is

this projection of the stochastic properties of the discovered deaths onto the undiscovered ones that provides the key leverage for estimation.[4] Unfortunately, it is certainly plausible that undiscovered deaths differ intrinsically from discovered ones. Indeed, this reflects a major challenge to the utility of the capture-recapture method. We discuss further below the challenge of detecting from observed data whether capture-recapture methods meet the necessary requirements to proceed effectively.

In Section 3, we discuss two strategies for addressing non-homogeneity of detection within lists, namely (i) stratification (that allows for the probabilities of detection to vary across known subgroups defined by age or geographic region, for example), and (ii) explicit modeling of how the probabilities of detection vary randomly within multiple lists (and thus not according to known subgroups). Each of these situations involves more complicated modeling of the mechanisms that generate the lists (that is, the processes that determine which casualties are reported on which lists), while still requiring fundamental assumptions regarding list selection properties that allow for valid inference regarding the estimate of the total number of deaths.

4. **Relationships between the Lists** The essence of the capture-recapture idea emerges most simply with just two lists ($k = 2$). In this case it is usually necessary to assume that there is statistical independence between the lists in the sense that the probability that a particular casualty appears on one list does not depend on whether or not it appears on the other, and vice-versa. This is not verifiable from the observed count data from the two lists. Less standard alternative approaches to two lists also require assumptions that cannot be checked using the available lists. Independence of the lists is closely related to homogeneity as we discuss further in the next section.[5]

[4] We note here that this strict homogeneity assumption is a sufficient condition for capture-recapture estimation to be effective. It is not necessary as even with two lists, for example, it is possible to accommodate heterogeneity in one list if there is perfect homogeneity in the other. In addition, it is possible to have selection heterogeneity in both lists and still produce unbiased capture-recapture estimates if the heterogeneity in one list intersects with that of the other in a mathematically elegant way; however, such advantageous circumstances are unlikely to occur in practice with list creation processes.

[5] This is most easily seen by considering a simple example: suppose casualties are equally split between two (unobserved) types, the first of which is always observed in each of two lists whereas the second type is only selected for each list 50% of the time on average. Then, if the two lists are assembled independently by random sampling that reflects these selection properties, the probability of a random casualty being on both lists is 5/8, whereas the probability of being on either list separately is ¾. Thus, appearance on the two lists is not independent given that type is unobserved. The dependence between the two lists occurs even though the lists were sampled

If our objective is to provide credible lower and/or upper bounds for the number of war dead, rather than a central estimate, then the weakness of having to assume independence in the two-list case is considerably attenuated because we can accommodate knowledge of inter-source dependencies into our bounds. In other words, an independence-based estimate can serve merely as a springboard for a discussion of possible list dependencies and their effects on the range of plausible war-death numbers. Providing clear and precise methods for determining bounds, and their associated uncertainty, would be a valuable topic for future research, particularly when there are many lists with potentially complex relationships so that pairwise dependencies are not easily described.

One of the major attractions of using more lists ($k > 2$) is that the assumption of list independence can be substantially weakened. However, it is always the case that some unverifiable assumption (typically that, at least, the k-order interaction is zero) is required to perform an analysis. (In general terms, the lack of a k-order interaction simply means that the intricate dependencies amongst any set of k-1 lists are not influenced by whether or not a death is on the remaining k^{th} list.) The nature of such an assumption becomes very hard to articulate to interested lay consumers of the information and, indeed, even to people with good statistical training, and thus hard to assess with regard to plausibility. This is an important problem in the context, for example, of a truth commission, one purpose of which is to provide explanation and closure to the families of victims. Thus, even when more than two lists are available we recommend initially calculating all possible two-list estimates (or bounds to be precise)--each assuming independence--accompanied by a discussion of possible dependencies. This can serve as a basis for understanding more complex estimates should the latter be required.

Section 3: How Strongly do Casualty Lists Deviate from Capture-Recapture Assumptions? Assumptions will be violated in any application of a statistical model. The interpretation of results requires judgment regarding the impact of these deviations from model requirements rather than merely noting their inevitable existence. This section provides

independently, with the association induced by the heterogeneity of selection probabilities across the two types that is correlated between the two lists. The cross tabulation of list counts does not provide sufficient information to distinguish between selection heterogeneity and dependence between the list construction processes.

some guidelines for making these assessments in applications of capture-recapture estimation to casualty estimation.

Just as unhappy families are unhappy in their own way (according to Tolstoy), each conflict and data-gathering effort is unique so that lists of war dead will violate the assumptions of capture-recapture analysis in their own distinct manner. Nevertheless, there are some general points we can make about the applications with which we are most familiar. We first address the quality of the available data before turning more directly to the assumptions themselves.

(a) Data Quality: Perfect Matching In many epidemiological applications, unique and reliable identifiers of individuals, such as social security numbers, are available. Individual information on people who have been killed in conflicts will often be much more limited and suffer from less validity, perhaps only using simple names and addresses at best, potentially incompletely supplemented by circumstances of death.[6] Each list must be cleansed of any spurious (i.e. false) casualties. One solution might be to attempt to validate all casualty records against population census lists or equivalent official data records including voter registries. However, even in cases where such records exist, this may still be difficult for vulnerable populations and children. Moreover, such validation work may require considerable resources and will be impossible in many situations.

When, information is collected retrospectively, with recall periods greater than a few months, then dates of deaths, or even the fact of whether or not a death has occurred, will often be recalled with considerable inaccuracy. Other details, such as victims' demographics or the circumstances of their deaths will often be incorrect. Any inaccuracies complicate the matching of deaths both within and between sources.

Much of the capture-recapture work in the casualty field has relied on the accuracy of extremely distant memories. For example, recall periods reach to nearly 40 years in Ball (1999, 2003), 20 years in Ball et al., (2003) and 25 years in Silva and Ball (2008). Much information supplied in this way must suffer from inaccuracies that are hard to detect, thus making it difficult to determine when different sources are reporting the same death.

[6] Even names might not be useful in matching in, for example, Sikh areas where normally men are named Singh and women are named Kaur.

Estimates of war deaths are necessarily inflated in capture-recapture estimates to the extent that the same deaths are recorded multiple times but not successfully matched within and across lists.

In addition to systematic errors caused by too little matching of identical deaths, there are also random errors in the matching process, an accounting of which must be built into the error bounds for a final count estimator if these are to be realistic. One method to incorporate random matching error is to mark all inconclusive matches with (subjective) matching probabilities as judged by coders. One then can perform a large number of random computer simulations in which records judged to match with probability of *p* will match in $100 \times p$ percent of the simulations. That is, on each computer run-through, all inconclusive matches are randomly resolved as matching or not based on their assessed matching probabilities, and then, after all these resolutions are complete, an appropriate algorithm calculates a capture-recapture estimate. The final estimate is then the average of these many estimates, each flowing from a particular resolution of all the uncertain matches. The final confidence interval accounts for the variation across these many estimates in addition to other forms of statistical error.

It is important to note the considerable literature on probabilistic record linkage that may usefully be applied here. For example, Fienberg and Manrique-Vallier (2009) discuss links between statistical ideas for record linkage and capture-recapture estimation, with data from casualties in Kosovo used for illustration. In addition, there is research on capture-recapture methods in situations without unique identifiers, a literature that formalizes the suggestion in the last paragraph (Laska et al., 2003; Caldwell et al., 2005). Nevertheless, there are thorny issues associated with matching, and the process is particularly complicated when the data are retrospective. For example, some individuals on some lists may be ‘anonymous’ in that their deaths are recorded accurately but without any identification whatsoever. Rules must be determined to handle such cases, and to account for additional estimation variability introduced by their presence.

Methods used to implement the matching should ultimately be entirely transparent and reproducible. Consumers of casualty estimates must be convinced of the validity of the matching effort since small systematic errors or biases may cause large differences in total count estimates.[7]

(b) Coverage and Homogeneity The most basic coverage and homogeneity assumptions require that each list employed in a capture-recapture estimation behave like a well-designed simple probability sample.[8] However, in practice, lists may reflect substantial "non-random" characteristics. There seems to be a common misperception that capture-recapture estimation can always transform "non-random" convenience data into unbiased statistical estimates complete with quantified sampling errors.[9] Of course, such a claim cannot always be true. As noted above, the assumptions underpinning any statistical model will never be satisfied in any kind of strict sense. However, practitioners must make a case that violations of the assumptions of the models they are applying are not so severe as to call the main results of an analysis into question. We recommend that an important part of the write-up of any capture-recapture estimate should include an argument that it is acceptable to treat the lists that are used in estimation as if they were well-behaved probability samples from appropriately modeled data-generating processes. We return to the topic of how to use statistical models to accommodate various list properties and relationships further below.

In seeking out multiple lists to understand the extent of casualties, it is of course natural and laudable to try to find new lists that document casualties that are likely to have been missed by other known lists. The acquisition of complementary lists allows a researcher to more fully cover all regions and types of casualties. In this sense, it is desirable that different lists 'fish in different ponds'. The problem is that when different lists focus on different types of victims an essential assumption for the most straightforward version of capture-recapture estimation is potentially violated, specifically that all individual deaths have equal chances of being

[7] There is a substantial amount of other literature on dealing with matching uncertainty including Lee (2001), Tancredi (2009, 2010), and da-Silva (2009).

[8] As noted earlier, and discussed in further detail below, these assumptions can be substantially weakened when there are multiple lists ($k > 2$).

[9] Landman (2006) writes "The key difference between the statistical estimation used in public opinion research and MSE [capture-recapture] is that where public opinion research uses random samples of the population, MSE uses multiple non-random samples of the population. Both forms of analysis produce statistical estimates with associated margins of error, ...". Material posted on the Benetech website (http://hrdag.org/resources/mult_systems_est.shtml) also seems to embody the misperception that capture-recapture necessarily transforms "convenience samples" into good statistical estimates.

listed on any particular source. The impact of performing capture-recapture estimates based on lists that focus on different types of victims can be very large indeed. Suppose, for example, that separate lists give accurate casualty lists, one for Shiites and one for Sunnis, with minimal overlap between the two. The combined list would provide a full and accurate count, whereas the simplest capture-recapture estimate based on the disaggregated counts would considerably overestimate the actual total count. In such cases, and when there are more than two lists, more complex statistical modeling will be required to accommodate the lack of homogeneity: we return below to two strategies for addressing violations of the strictest version of the homogeneity assumption.

There appears to be a misperception that capture-recapture estimates can always uncover deaths—at least, in aggregate--that differ systematically from deaths that do appear on lists.[10] The problem is that, when the coverage and homogeneity assumptions are satisfied, then deaths estimated as missing in all the sources must necessarily inherit the (possibly complex) sampling characteristics of the known deaths. That is, the capture-recapture assumptions require that it is only due to pure chance that deaths fail to appear on lists. Thus, estimated but unlisted deaths cannot differ fundamentally from listed ones, although it is a challenge to convey in full generality the precise meaning of this fact.[11] The concept is simplest when there are only two independent and perfectly homogeneous lists: in this case if, for example, ten percent of the victims on a particular list are females then, aside from sampling error, 10% of the victims not listed on that source should also be females.[12] The same should be true for any characteristic of the victims such as ages, or the identities of perpetrator groups who killed them. Unfortunately, the situation rapidly becomes more complicated when

[10] For example, "the purpose of multiple systems estimation is to determine, given a number of independent sources of information, (comma added for clarity) what might be missing from them. The reason one would do this is that the information that is missing could be in some way systematically different from the . . . data which is known." Ball, P., Testimony before Kosovo trial, p. 10223.

[11] Capture-recapture methodology has to make an untestable act of faith that in some respect missing individuals resemble listed ones . . . It may be that the value of multiple lists lies less in the provision of a numerical estimate of population size, more in identifying when substantial undercounting exists, and prompting investigation of possible causes." (Cormack et al., 2000). Bishop et al. (1975, p. 254) also note that the capture-recapture approach can sometimes be misleading "since we are assuming that the model which describes the observed data also describes the count of the unobserved individuals. We have no way to check this assumption."

[12] In fact, approximately 10% of the victims both on and off each list used in the estimation should be female if the coverage and simplest homogeneity assumptions of capture-recapture estimation are satisfied.

there are heterogeneous list detection probabilities and/or more than two lists even though it is still possible to proceed in some such situations even though demographic characteristics of list members may vary substantially across the lists. In more complex scenarios, it is not always clear how to distinguish easily between cases where appropriate modeling can be effective (in dealing with the complexity of the list properties and their inter-relationships) from those where it cannot.

To produce a convincing capture-recapture analysis it is important to use the available lists to assess the assumed coverage/homogeneity properties to the greatest extent possible. Each list can be examined for time trends, demographic characteristics of victims, perpetrator groups and any other major covariate for which there are data. As noted, perfect homogeneity assumes that each list can be viewed as a random sample from the *same* (unknown) master list of total casualties, so the sample distributions for each covariate should be similar across lists subject to sampling error. In other words, each source should show similar time trends, a similar geographical distribution of deaths, similar victim demographics and other breakdowns for which data are. Consistency of covariate distributions across sources cannot prove perfect homogeneity, but strong inconsistencies will demonstrate possible heterogeneity and the need for more complex estimation techniques at the very least. A systematic approach to this exercise for Sierra Leone is described in Gohdes (2010), where three data sources paint very different pictures of the conflict.

Of course, the strictest homogeneity assumption is quite likely to be violated in the context of casualty lists since it will rarely be true that available lists of deaths really will be straightforward homogeneous random samples from the complete list of conflict dead. In the canonical capture-recapture application of estimating the number of animals on some territory, researchers often have the luxury of collecting designed random samples so as to make this homogeneity assumption plausible. In the casualty setting, however, the major tools of random sampling are often denied the investigator. The lists tend to arise in a happenstance fashion and/or are targeted to collect a specific kind of casualty such as victims of the government or from a particular religious group.

As noted earlier, with multiple lists ($k > 2$), it is possible to proceed with estimation even if homogeneity is not satisfied, by exploiting various statistical modeling approaches.[13] However, stratification remains a plausible first response to the problem of heterogeneous capture probabilities.[14] Effectively this approach disaggregates the estimation problem into 'smaller' problems in fixed covariate subgroups. For example, it will almost always be essential to consider the age of a victim as a potential stratification factor since it is likely that child deaths may be underreported in some lists due to the greater social visibility of adults.

In principle, stratification can help because it is possible that within-list capture probabilities are heterogeneous at the aggregate level but can still be usefully treated as homogeneous within strata such as geographical areas or victims of particular perpetrators.[15] Ball et al. (2003) works with two sources that mainly record victims of the government in the Peruvian conflict together with a third source that records both victims of the government and the guerrillas (Landman and Godhes, 2011). These characteristics of the lists led Ball et al. (2003) to stratify by perpetrator in their analysis as a means of addressing heterogeneity.

In most cases stratification may not produce sufficient homogeneity to underpin good capture-recapture estimates. Sometimes the necessary grouping variable--upon which a good stratification scheme could be based--may simply be unknown or not measured. In such cases, multiple lists ($k > 2$) can potentially allow the statistician to address the failure of the strictest assumptions through use of more complex modeling and estimation methods that account for the heterogeneity (or induced dependencies between the lists).[16] One is however faced with the challenge of using available concomitant (e.g. covariate) data to distinguish between situations where modeling is adequate to address unknown complex sampling characteristics and those where such modeling is not up to this task. We are not currently aware of simple and general direct diagnostic methods that can support the use of statistical models addressing complex data-generating mechanisms with the same

[13] Usually, the necessary estimation procedures are couched in terms of log-linear models applied to cross tabulation of casualties across the available lists (Bishop et al., 1975).

[14] See, for example, the work in Guatemala, Peru, and Colombia (Ball, 1999; Ball et al., 2003; Lum et al., 2010).

[15] In the general case where statistical modeling can potentially address complex heterogeneity, it may still be easier to support the modeling within strata as preferable to aggregate modeling.

[16] Again, this is usually approached through log-linear modeling.

level of confidence as is the case for the simplest situation with perfectly homogeneous/independent lists.

Note that complex modeling within the context of capture-recapture estimation differs essentially from standard statistical approaches to complex sampling techniques that allow for varying selection probabilities. In principle, the degree of complexity of a designed sample should not matter so long as one can access a sufficiently sophisticated statistician who can make valid estimates based on knowledge of the sampling design. If, on the other hand, we have casualty lists that are generated by uncertain or unknown methods then we can still proceed as if the lists were generated by some known random mechanism. However, the problem now is that we will probably have only limited possibilities for actually understanding and validating this data-generating mechanism. In this case, complex sampling does not translate simply into a challenging puzzle for a statistician—rather, it requires us to make complicated assumptions before we can proceed to capture-recapture estimation. In short, with a designed sample, the complexity of the sampling scheme is a known fact and this complexity is built into the estimation, whereas in capture-recapture estimation the nature of the sampling scheme is an imposed assumption that necessarily influences the results. One hopes that more lists will reduce bias stemming from incorrect assumptions, although it would be good if we could quantify this expectation since, in general, it is not desirable to make complicated assumptions when there is little information.

In summary, we recommend that stratification strategies in a capture-recapture estimation should include at least four basic components. First, there should be an analysis of how the lists that are being used have been constructed to identify likely sources of heterogeneities in coverage such as uneven emphasis on certain time periods, geographical areas or types of victims. Second, there should be an analysis of covariate information at the macro level. This can further expose major sources of heterogeneity that might be possible to address through stratification. Third, based on the first two points, stratification schemes should clearly address the heterogeneity problems that have been identified. In the literature we find some, but not enough, motivation for rather elaborate stratification schemes that have been used. Finally, extending the above discussion, researchers should assess heterogeneity within each stratum where possible, and either make a case that departures from homogeneity within strata are not likely to be so pronounced as to undermine the final estimates or that these heterogeneities can be modeled successfully.

Researchers trying to address heterogeneity through stratification should also bear in mind the tension between increased bias at higher levels of aggregation versus greater variability at lower levels due to smaller strata sample sizes. It might make sense in some cases where different lists are focusing strongly on different types of victims to simply combine the lists.[17] Further research on this topic might provide useful practical guidance on how best to proceed.

It would be remiss of us not to note that there are serious efforts in the capture-recapture literature to model heterogeneity using other approaches than log-linear models applied to the cross tabulation of casualty counts by lists (Manrique-Vallier, et al., 2011, also touches on this topic). Many of these approaches involve a form of mixing assumption where it is assumed that the probability of (a specific) list detection varies across individuals and that this can be described usefully. It is possible that this variation itself might vary across lists, and across time (see, for example, Coull and Agresti, 1999). In many cases, these ideas invoke a notion of what might be called “listability” (a latent class variable) that randomly varies across individuals but is unobserved.[18] In essence, such approaches involve assuming some structure for the population distributions of listability.

(c) Interdependence of Lists Although the assumptions of homogeneity and list dependency appear to be quite distinct, these are, in fact, essentially equivalent concepts expressed differently (Hook and Regal, 1995, pp. 255-256), as noted earlier. Formally, the bias of the naïve estimator can be expressed in terms of a contribution completely from dependency, alternatively from unequal capture probabilities completely, or from both sources together. Thus, much of the following discussion returns to issues previously discussed in terms of homogeneity assumptions.

As with homogeneity, assumptions regarding the independence of lists may be made more plausible with designed samples, as in applications to estimating wildlife populations. This is again not generally possible with casualty counts. For example, it is likely that casualty lists tend to borrow

[17] We return briefly to this point in Section 4.

[18] An alternative mixture approach to casualty counts was suggested by Manrique-Vallier and Fienberg (2008) using latent classes where individual casualties are allowed to belong partially to all classes, known in the broader statistical literature as a ‘Grade of Membership’ model.

from each other during their creation and updating, a characteristic that violates independence instantly. Comparison of total count estimates based on different pairs of lists can sometimes identify pairwise dependence. More subtle forms of dependence that violate assumptions are harder to detect, as previously discussed.

As noted, the independence assumption is essentially required in two-list estimation, but can be very much relaxed when multiple lists are available, mirroring our discussion of homogeneity.[19] While it is true that dependency assumptions are less restrictive with more than two lists, one still must make some kind of assumption ruling out more complicated types of dependencies between lists (e.g. the dependency between lists A and B is not influenced by whether a casualty is on list C or not). In addition, there is the usual statistical price of less precision--implying wider confidence intervals in the ultimate total estimate--to the extent that complicated models are required to allow for more and more complex types of dependencies between lists.

An additional problem is that with more than two lists it becomes a challenge to explain the nature of the interdependency assumptions that are being invoked, particularly to non-technical audiences, and to assess whether all important dependencies can be suitably captured by a statistical model. This again raises the question of how to effectively use concomitant information to support the use of a particular statistical model

This need for simplicity and transparency suggests the possibility of focusing—initially at least—on two-list estimation using all different combinations of two lists, accompanied by discussions of heterogeneities and dependencies, that then feed a discussion of likely lower and upper bounds for total war deaths.[20] As a simplistic illustration, we use the three lists of Peru data published in Ball et al. (2003) to obtain three two-list capture-recapture estimates (after stratifying by perpetrator) of 14,000, 56,000 and 99,000. These large differences in pairwise capture-recapture estimates immediately suggest some list dependence. The key issue, of course, is whether they reflect only pairwise dependence so that a three-list

[19] Usually, the necessary estimation procedures are couched in terms of log-linear models applied to cross tabulation of casualties across the available lists (Bishop et al., 1975).

[20] In doing so, we emphasize that that no single estimate based on two lists solely is likely to be defensible or reliable taken on its own, even if only used to create an upper or lower bound. The combination of all possible two list estimates has to be considered in their entirety, and in light of external knowledge about list construction and estimates based on more complex modeling.

estimator will be effective. As noted, it is not clear how to use any available covariate data to assess this assumption in any meaningful way, thereby provoking a healthy skepticism about any three-way estimate derived from the same data. Ball et al. (2003) do use three-list estimation with stratification both by perpetrator and geography to arrive at an estimate of 69,280 deaths with a 95% confidence interval of 61,007 to 77,552. This is considerably higher than the approximately 24,000 deaths plus disappearances recorded by the Peruvian Truth and Reconciliation Commission, so that this has to be considered a rather bold estimate in the first place. It would clearly be helpful here to provide a simple explanation of the strong deviation from the lowest two-list estimate.

Section 4: The Analysis and Interpretation of Capture-Recapture Estimates of War Deaths Even if the assumptions underlying capture-recapture estimation are reasonably well satisfied, there remain a number of thorny issues to address, including (i) an overabundance of possible models to choose from when there are many lists, (ii) accounting for all sources of error in assessing the accuracy of casualty estimates, and (iii) conveying the nature of the estimates to an educated public.

Model Selection It is important to understand that, in general, there is not a single capture-recapture model[21] but, rather, a large number of capture-recapture models with the number of available models growing substantially as the number of lists grows.[22] Moreover, different plausible models can yield extremely different estimates, and there is no clearly validated method for choosing among them. This means that the uncertainty affecting capture-recapture estimates based on more than two lists is greater than is generally understood with most confidence intervals that have been published in the literature likely understating this uncertainty.

With two lists, we must select, in principal, from among only six models although really only one of these is useful in practice (the model that allows the probability of capture to differ between the lists but assumes list independence, as discussed earlier). With more lists, the possible model choices grow rapidly so that with four lists there are potentially more than

[21] For simplicity here, our remarks refer to the use of log-linear models although the issues are broadly applicable to all modeling strategies.
[22] Of course, an advantage of the expanded choice of models is the potential ability to accommodate more complexity in list relationships and heterogeneity in list selection probabilities.

32,000 log-linear models from which to choose (32,766 to be precise).[23] This large number of models can be reduced substantially by restricting the kind of models to be considered plausible,[24] and by additional restrictions such as assuming that all dependencies higher than a certain order do not exist. The point here is not to focus on the exact number of appropriate models to assess, just that this number grows quickly as the number of available lists increases. This occurs because of the number of possible list dependencies that can be included or not. For example, with only three lists—with three pairwise dependencies possible--there are seven possible ways to include these effects: allow for all three pairwise dependencies, only include two of them (there are three different ways to achieve this depending on which pair of lists is assumed independent), and only include one of them (there are three different ways to do this depending on which pair of lists is assumed dependent). This plethora of possibilities only gets more daunting when we have four or even more lists available.

There are several potential problems that arise from having to select from an abundance of models. First, it is possible that total casualty estimates will vary strongly between models, even among ones that apparently describe the observed data effectively. This is illustrated in Lum et al. (2010) where the two most plausible models in estimating the total number of casualties in Casanare differ by a factor of three despite both fitting the data almost equally well. The problem deepens when one realizes that there appears to be little or no systematic validated technique for choosing amongst such models. Averaging over a set of possible models is one way to try to approach uncertainty over what the best model is, reflecting the statistical premise that averaging a set of uncertain estimates is likely to be more reliable than choosing any specific one. York and Madigan (1992) and Madigan and York (1997) discuss averaging across models in a Bayesian framework, and an application of this type of approach to the Casanare data is given in Lum et al. (2010).

Second, confidence intervals covering the final estimate are usually based on the assumption that the selected model is known with certainty to be correct. This is, of course, a problem in many applications of statistical

[23] Lum et al. (2010) use 15 different casualty lists in Casanare, allowing potential consideration of 2 raised to the $(2^{15}-1)^{nd}$ power number of models (less 2 if we always exclude the log linear model with only the intercept and the saturated model which is over-determined as previously noted)--this formula corrects a minor typographical error found there.

[24] For example, by only considering hierarchical log-linear models.

modeling. The point is that the uncertainty associated with model selection is not accounted for in many reported confidence intervals so the published margins of uncertainty are too small. We return to this below in our comments on error estimates. At the very least, it is valuable that investigators report as many of potential model estimates as possible so that consumers can see the level of empirical variation associated with various model choices. A graph akin to Figure 1 in Lum et al. (2010) that shows various model estimates, plotted with a measure of how well they fit the data, is a good start. Such creative approaches may be necessary since with more than a few lists it will be impractical to report all available estimates in a table.

How many Lists? We have previously noted immediate difficulties with the strict assumptions necessary to allow two lists to provide reasonable estimates. Three lists are better than two, at least in the sense that the independence assumption can be weakened. Does this inevitably mean that 10 lists are better than three? From an ideal statistical viewpoint this should still be true, and twenty lists should be even better than 10. But this assumes that all lists are equally valid and useful, and this is extremely unlikely to be true in conflict situations. What is the 'right' number of lists that can support a reliable estimate? Should certain lists be combined? If so when and under what circumstances? These are all important practical questions that require statistical input and guidance.

A potential first approach with several lists is to focus mostly on multiple two source estimates using the largest sources, where one can predict the likely dependence (positive or negative) across any two sources using external information and/or information from statistical models. The derived estimates would then be known to likely under or overestimate the true count. With several such analyses one might derive a likely lower or upper boundary (and uncertainty estimates on these bounds). At the very least, this kind of analysis may underpin more complex estimates making them more readily transparent. However, in a detailed but small example, Cormack et al. (2000) found that—paradoxically—discarding the list with the highest coverage was necessary to achieve a satisfactory estimate and inference, in turn confirming that sometimes it may be better to work with a subset of available lists. They were able to identify this since a gold standard total estimate was available; the challenge, of course, is to determine an appropriate strategy in the absence of any external validation whatsoever. It may also be better to ignore some sources rather than pooling them because the relationship between a pooled constructed list and some other source may not be as readily predictable as the relationships between pure sources may be.

It is simplistic to assume that all lists should be assumed to be of equal validity. We have already noted that some lists tend to borrow heavily from each other so that it might be best to combine such lists rather than trying to disentangle the dependencies. This suggests that a deep understanding of the ways in which the lists are generated may be as important as trying to rely on intricate statistical models to adjust for all forms of association across lists. At some point, the marginal value of an additional list may be small in terms of improving accuracy and might even increase bias if the list is poorly constructed or intricately connected with all other sources. It would be desirable to have a formal statistical procedure to decide when to pool lists and when not to attempt to add a list. A general policy for list selection and combination that is widely effective remains an open question.

Accounting for Error (confidence intervals) Casualty estimates should be accompanied by some assessment of possible error, something that would at least take the form of a 95% confidence interval in a traditional estimate. Interpreting confidence intervals from capture-recapture estimates on the basis of repeated experimentation (frequentist inference) is somewhat problematic since it is difficult to consider an appropriate (and approximate) stochastic mechanism that truly generates the available lists. Proper model-based confidence intervals should account at least for variation in estimates due to random sources of error in the discovery and de-duplication of deaths, errors in matching deaths across lists, and model selection. Currently, capture-recapture confidence intervals in the casualty estimation literature generally account only for sampling error in the discovery of deaths and not other error sources.

An alternative approach to exploring a fuller range of error sources is to simulate the impact of the various error sources on a computer, a procedure formally known as ‘bootstrapping’. In principle, these simulations can be made reasonably transparent through the use of devices such as flow charts that explain the computer procedures. In practice, however, realistic simulations of errors in capture-recapture estimates raise complicated issues that extend beyond the scope of the present paper.[25]

Bayesian approaches provide an alternative approach to interval estimation and have been widely studied in the context of capture-recapture methods (see Smith, 1988; Madigan and York, 1997; Ghosh and Norris, 2005; Lee et al., 2003, for example). These methods usually require specification of an assumed prior distribution of the casualty count; subsequently, a posterior distribution of the count is obtained conditional on the observed data from the lists using Bayes rule. This approach has now been applied to casualty applications (Lum et al., 2010), where, as we noted above, Bayesian averaging across models is used to combine a large number of estimates. Posterior means can be used as point estimates with credible intervals derived from the full posterior distribution.

In addition to accounting for random variation in the way lists are created, matched and analyzed, it is also important to account for systematic error associated with potential violation of assumptions and model misspecification. While some of this is addressed by averaging estimates across different models and assumptions, some form of sensitivity analyses may also be desirable. This largely boils down to sharing with readers a wide range of estimates over a variety of plausible models and assumptions, rather than just presenting one final model or one final estimate that averages over many unseen models. Confidence intervals can thus be extended into some form of plausibility intervals that account for both sampling and systematic errors. Achieving consistent analytic and reporting requirements is an important goal.

Transparency of the Method and Assumptions We noted earlier that the availability of several lists allows not only more data and therefore improved absolute lower bounds but also means that the methods make less demanding assumptions on list selection properties. On the other hand, even with as few as three lists, the assumptions underlying a proposed final count estimate are difficult to grasp. Without some level of satisfactory transparency, there is a serious risk of losing credibility. And without credibility, the whole point of an improved count is lost.

[25] See, for example, Buckland and Garthwaite (1991), Coull and Agresti (1999) and Amoros et al. (2008).

Unfortunately, accommodating deviations from implausible model assumptions, such as homogeneity, and errors in matching, requires an increasing level of complexity in statistical techniques, acting against the need to be transparent. Clear description of model, model selection, error assessment and sensitivity analyses should be reported with every analysis, although it is hard to see how these aspects can be fully assessed by anyone other than statistical experts.

Validation of Capture-Recapture Techniques in Casualty Estimation

Four basic approaches have been developed to provide casualty counts and estimates: (i) survey methods, (ii) capture-recapture estimation (iii) direct and indirect contemporaneous counts, and (iv) census and other demographic techniques. It is clearly helpful when more than one method is available for a specific conflict assessment, particularly when the available methods all lead to similar results. It can be just as informative, however, when methods yield divergent results in that it highlights the need to examine the assumptions and data on which each method depends, and determine and evaluate the conditions that may lead some approaches to questionable results. It is crucial therefore to look at comparisons of the spectrum of techniques as much as possible.

Currently, the situation where the most direct comparisons are available is Kosovo where there are three principal sources of casualty information: (i) the post-conflict survey of Spiegel and Salama (2000) (that is, method (i) above), (ii) the capture-recapture estimates from the AAAS report by Ball et al. (2002) (method (ii)), and (iii) the more recent and less publicized work of the Humanitarian Law Center called the Kosovo Memory Book (method (iii)).[26] The latter is an attempt to provide an exhaustive list of all Kosovo victims from the relevant time period, with some basic information provided for each victim.

There is insufficient space here to do justice to a full comparison of the results of each of these methods in Kosovo, an analysis that we intend to provide elsewhere (Spagat et al., 2011). In brief, there is considerable consistency between the three approaches. It is hardest to compare the results of the two other approaches with the estimate of Ball et al. (2002) since the latter covers a much shorter time period (albeit the most violent months), and does not provide estimated counts by age. For the time period

[26] www.hlc-rdc.org/index.php?lid=en&show=kosovo&action=search&str_stanje=1

March-June 1999, Ball et al. (2002) estimates 10,356 deaths with a 95% confidence interval of 9,002 to 12,122 (based on 4,400 unique named deaths from 4 lists). For this same period, the Humanitarian Law Center list yields 9,030 'murders' with a further 1,200 'missing.' The two methods provide similar quantitative results even if we assume that all missing people are still alive. The Spiegel and Salama (2000) survey estimates 12,000 deaths due to "war-related trauma" between February 1998 and June 1999 with a 95% confidence interval of 5,500 to 18,300. This compares to counts by the Humanitarian Law Center list of 11,401 'murders' and 1,567 'missing' for the same time period. These two sources also confirm a general age distribution pattern with the elderly facing a far higher risk of death than young adults. In addition, the monthly time series for the two methods track each other generally although, of course, there are far larger margin of errors associated with these disaggregated data. In summary, all three methods emerge with credit and validate the other techniques in turn. Overall, this example appears to be a particularly favorable case for capture-recapture estimation in that the different lists used appear to be quite consistent with each other except in locality, and heterogeneity in the latter was also accommodated through geographical stratification. This suggests that the necessary assumptions and selected models are reasonable in this case.

Other opportunities for comparison of the alternative estimation techniques occur in Peru where Ball et al. (2003) contrast a capture-recapture casualty estimate for the Department of Ayacucho with demographic calculations of the amount of excess mortality using national census data from 1981 and 1993 (Ayacucho being the Department most affected by armed internal conflict). A description of the comparison is available in Appendix 2 of Ball et al. (2003), indicating that the excess deaths estimate exceeds the capture-recapture estimate of violent deaths by 30%, taken there as tending to confirm the capture-recapture estimate. However, the general similarity of the estimates is not particularly compelling because the ratio of excess deaths to violent deaths ranges between 1 and 17 in 13 conflicts assembled in the Geneva Declaration (2008, p. 40) so that the specific analysis serves principally to illustrate a generally accepted pattern.

Finally, Silva and Ball (2008) discuss similar results obtained from a retrospective mortality survey (16,000 deaths plus or minus 4,400) in Timor Leste with that obtained from using a two-list capture-recapture estimate

(18,600 deaths plus or minus 1,000), providing some mutual validation for the two techniques.

Section 5: Challenges for the Future

Additional Data Currently capture-recapture estimation for casualty counts has essentially adapted techniques from other applications in wildlife, ecology and public health. However, casualty lists do not reflect capture-recapture in their creation and are often constructed contemporaneously. While we have noted several issues related to this distinction as they relate to assumptions, it is possible that attempts should be made to collect and exploit additional information pertaining to the development of casualty lists. A key feature of some casualty applications is that being named on a list reflects an event in time separate from when the casualty occurred. As yet, we have not seen this information widely exploited, recognizing that there will be inevitable errors in reporting dates or substantial levels of missing information.[27] We note that the time from an event until a casualty is recorded is likely a random variable whose distribution varies from list to list, and even within a specific list. The fact that this distribution is not consistent and is unlikely to be known *a priori* means that techniques from infectious disease incidence counting—like back-calculation—cannot be implemented here. Nevertheless reporting delays might usefully be recorded when possible and used to characterize lists at the very least.

Simulation and Test-bed Methods In infectious disease research, both deterministic and stochastic explanatory models based on complex systems of differential equations have played a major role in exploring the properties of epidemics and interventions to change their course. Such models provide the basis for effective simulations that allow an investigator to quantify the sensitivity of estimated incidence patterns to assumptions and input parameter assumptions, and to compare various approaches to intervention. With casualty reporting and counting, it would be extraordinarily helpful to have analogues of such models to compare estimation techniques in cases where the true processes and counts are known. In addition, 'test-bed' data sets created from such models may

[27] The Iraq Body Count records this delay and has used the information to address the incompleteness of their counts in that if a great many casualties were missing from their list, occasionally the media would uncover casualties some time after the deaths occurred; in fact, virtually 100% of all recorded car bombings were reported to the media within 24 hours, for example.

provide valuable insights into the sensitivity of estimation methods to particular assumptions or characteristics of how the data is imperfectly generated from underlying events. In developing such models, perhaps using ideas from agent based modeling (Bonabeau, 2002), it is necessary to (i) simulate the development of a conflict and various sources of casualties and (ii) simulate the various methods of data collection that reflect the vagaries of discovery/reporting of these casualties. The processes generating casualty events may or may not be linked to those generating the data/lists. Simulations based on these models would permit various forms of sensitivity analyses. The act of creating such explanatory models in itself raises interesting questions in that, in some explanatory models, rapid increases in casualties may either tend to be sustained or lead to sudden drops in counts (in some cases, the deaths of certain key individuals cause a decline in further deaths). For further provocative reading in this regard, see Epstein (2006) and Williams (2007).

Data and Software For statistical methods to be fully developed and validated in the area of casualty counting, two additional factors are important. First, there is need for existing data sets to be made widely available, allowing for redaction and protection of individual identification where necessary. While this may not allow other investigators to fully recreate the issues involved in matching lists etc., it will provide opportunities to validate various estimation approaches, strategies for model selection, and error estimation and sensitivity analyses. The availability of casualty data from Kosovo, Sierra Leone, Timor-Leste, Liberia and Casanare from links on the Benetech data page[28] is an important step in this direction. While summary data is of value, the more detail that can be provided the better, as it allows for more nuanced validation of methods.

On a related issue, it is imperative that available open source software be made available to apply capture-recapture estimation in the casualty setting. Within R, the Rcapture package provides routines for capture-recapture estimation based on the log-linear modeling approach (Baillargfeon and Rivest, 2007), although it would be helpful to have the methods and documentation more directly related to casualty estimation to allow for wider use in this setting.

[28] www.hrdag.org/resources/data_software.shtml

Section 6: Recommendations Finally we provide a simple list of practical suggestions for future capture-recapture estimates of war deaths.

1. Think about the purpose of estimating war deaths in the first place and whether it might be preferable to simply give a good lower, and perhaps an upper, bound on the number of war deaths. If, for example, the purpose is to support legal prosecutions, then well documented lower bounds are probably more valuable than count estimates.
2. Presentations of capture-recapture estimates must make a case that it is reasonable to treat lists as amenable to modeling as well-behaved probability samples with appropriate assumptions to address observed or known list heterogeneity and dependence.
3. If stratification is meant to address departures from random selection of deaths onto lists (heterogeneity), then the stratification scheme must be based on the appropriate sources of heterogeneity. Again, as good a case as possible must be made that, within each stratum the cross-tabulated list counts can be modeled appropriately.
4. Gather as much information as possible on how each list has been constructed and provide readers with a detailed write-up on the nature of each list.
5. Consider the likely dependencies among lists, i.e., the extent to which the appearance of a death on one list raises or lowers the probability that this death will appear on other lists. Analyze the impact these dependencies are likely to have on estimates.
6. Publish a list of two-list estimates for all pairs of lists (three estimates with three lists, six with four lists etc.), using appropriate stratification if necessary. These will help to identify both dependencies between lists and departures from the simplest homogeneity assumptions, and may be used to establish initial lower, and possibly upper, bounds. Considerable emphasis should be placed on these two-way estimates because they are much easier to understand and interpret than are estimates based on three or more lists.[29] This analysis will also aid in the interpretation of more complex estimates.
7. Make matching as transparent as possible. At a minimum provide a large random sample of several categories of matches and non-

[29] The availability of a third list allows, of course, an empirical assessment of pairwise list dependency that may support external knowledge of list relationships. This is more complicated when four or more lists are available but still worth pursuing in adding insight into the validity of any specific two list estimate.

matches, e.g., matches considered definitive, matches considered likely, etc..

Section 7: Discussion Statistical ideas and estimation methods have provided extraordinary tools for various applications in the social sciences. Having said this, not every statistical tool can be implemented immediately for a specific application. The recent past has seen a burgeoning use of capture-recapture estimation as part of the development of a science of casualty counting. The assumptions are certainly open to scrutiny in general and in each specific application. While it is true that the assumptions underpinning any statistical model are never satisfied in a strict sense, it is important in this sensitive area that there be a reasonable alignment of theory and reality. Given the intense political and media interest in reported casualty estimates, it is crucial that statisticians agree that any assumptions used are robust to plausible violations. Estimates must be reported with full acknowledgement of the shortcomings or difficulties associated with uncertainty, a challenge that is difficult to translate into the political arena or standard media reporting. It is important to be aware of the tension between the need for statistical rigor and accuracy and the particular uses that are made of estimates. In many situations, it may be best to never report a single number or point estimate.

Inevitably, the marginal value of a capture-recapture estimate is lessened when other counts or estimates are available to validate the capture-recapture results since this essentially means that the same information is available from other sources. Even so, the scientific importance of validation should not be underestimated. When there are only capture-recapture estimates, these are potentially valuable but it is unclear how much weight should be placed on them without validation by other approaches. It is similarly unclear how much we can depend on capture-recapture estimates when these disagree widely with estimates based on other methods. Certainly, capture-recapture estimates that greatly exceed the counts of documented deaths from the combined underlying lists should be scrutinized with particular care. Ironically, estimates that differ strongly from documented counts are precisely the cases where capture-recapture estimation potentially has the most to offer.

Many casualty lists are created by parties with personal or political views that may influence the way lists are created, manipulated and analyzed. This is of particular concern since decisions may be taken early in a data-collection process that bias the results but might not be known to independent statistical investigators. It is a challenge to deal with the role of personal convictions in any statistical analysis, and particularly the estimation and reporting of casualty counts. Wide availability of the data, transparency in all decisions relating to the lists' creation, matching, and statistical analysis is a fundamental requirement for credibility. To the greatest extent possible, casualty data should be analyzed by independent statistical investigators.

We reiterate that casualty counting fundamentally differs from many traditional applications of capture-recapture in that the lists are far from designed probability samples of the total number of casualties. As such, all such estimates should be treated with an appropriate level of skepticism. However, it is inevitable that people will keep creating such lists during conflicts for a variety of valid reasons, with such lists then exploited to provide more complete estimates in multiple ways. In addition, other methods for counting casualties also face serious statistical challenges and there is no method that is clearly best for all circumstances. A determination to count casualties pushes us to search for approaches that seek to both minimize assumptions and data problems and maximize validity, rather than attempting to generate perfect numbers.

Acknowledgements

The authors thank the following individuals for comments on earlier drafts of this chapter: Patrick Ball, David Banks, Josh Dougherty, Baruch Fischoff, Ernest Hook, Daniel Manrique-Vallier, Megan Price, Nigel Quinney, and Romesh Silva. Their feedback was enormously helpful and we are deeply appreciative; of course, all remaining errors are our responsibility.

References

Aaron, D. J., Chang, Y-F., Markovic, N., and LaPorte, R. E., "Estimating the lesbian population: a capture-recapture approach," *J Epidemiol Community Health,* 2003, 57, 207-209.

Amoros, E., Martin, J. L., Lafont, S., and Laumon, B., "Actual incidences of road casualties, and their injury severity, modeled from police and hospital data, France," *Eur. J. Public Health*, 2008, 18, 360-365.

Andreas, P. and Greenhill, K. M., "Introduction: The politics of numbers," In *Sex, Drugs, and Body Counts. The Politics of Numbers in Global Crime and Conflict.* Andreas, P., Greenhill, K. M., Eds., 2010, 1-22, Cornell University Press, Ithaca, New York.

Baillargeon, S. and Rivest, L.-P., "Rcapture: Loglinear models for capture-recapture in R," *Journal of Statistical Software*, 2007, 19, 1-31.

Ball, P., "Making the Case: Investigating Large-Scale Human Rights Violations Using Information Systems and Data Analysis," Report for the Guatemalan Commission for Historical Clarification, 1999. http://shr.aaas.org/mtc/chap11.html

Ball, P., "Using multiple system estimation to assess mass human rights violations: The cases of political killings in Guatemala and Kosovo," *Proceedings of the International Statistical Institute*, 2003, Berlin.

Ball, P. and Asher, J., "Statistics and Slobodan: using data analysis and statistics in the war crimes trial of President Milošević," *Chance*, 2002, 15, 17-24.

Ball, P., Asher, J., Sulmont, D., and Manrique, D. How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000, AAAS. Report to the Peruvian Truth and Reconciliation Commission (CVR), 2003. Also published as Anexo 2 (Anexo Estadı´stico) of CVR Report.

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and Asher, J., "Killings and refugee flow in Kosovo March-June 1999: a report to the international criminal tribunal for the former Yugoslavia," 2002, AAAS, Washington, DC.

Bishop, Y.M., Fienberg, S. E., and Holland, P. H., *Discrete Multivariate Analysis: Theory and Practice*, 1975, MIT Press, Cambridge.

Böhning, D., "Editorial – Recent developments in capture-recapture methods and their applications," *Biometrical Journal*, 2008, 6, 954-956.

Böhning, D. and van der Heijden, P. G. M., "A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations," *Annals of Applied Statistics*, 2009, 3, 595-610.

Bonabeau, E., "Agent-based modeling: methods and techniques for simulating human systems, " *PNAS*, 2002, 99, 7280-7287.

Brunborg, H., Lyngstad, T. H., and Urdal, H., "Accounting for genocide: How many were killed in Srebrenics?" Chapter 9 in *The Demography of Armed Conflict* Brunborg, H., Tabeau, E. and Urdal, H., Eds., 2006, Springer, Dordrecht, The Netherlands.

Buckland, S. T. and Garthwaite, P. H., "Quantifying precision of mark-recapture estimates using the bootstrap and related methods," *Biometrics*, 1991, 47, 255-268.

Caldwell, B. L., Smith, P. J., and Baughman, A. L., "Methods for capture-recapture analysis when cases lack personal identifiers," *Statistics in Medicine*, 2005, 24, 2041-2051.

Chao, A., P. Tsay, S.-H. Lin,W.-Y. Shau, and D.-Y. Chao (2001): "The applications of capture-recapture models to epidemiological data," *Statistics in Medicine*, 20, 3123–3157.

Cormack, R. M., Chang, Y-F., and Smith, G. S., "Estimating deaths from industrial injury by capture-recapture: a cautionary tale," *Int J Epid*, 2000, 29, 1053-1059.

Coull, B. A. and Agresti, A., "The use of mixed logit models to reflect heterogeneity in capture-recapture studies," *Biometrics*, 1999, 55, 294-301.

Da-Silva, C.Q., "Bayesian analysis to correct false-negative errors in capture-recapture photo-ID abundance estimates," *Brazilian J. Probab. Stat*., 2009, 23, 36-48.

Desenclos, J.C. and Hubert, B., "Limitations to the universal use of capture-recapture methods," *Int J Epid*, 1994, 23, 1322-1323.

Epstein, J. M., *Generative Social Science: Studies in Agent-Based*

Computational Modeling, 2006, Princeton Studies in Complexity Series, Prinecton University Press, Princeton, New Jersey.

Fienberg, S. E. and Manrique-Vallier, D., "Integrated methodology for multiple systems estimation and record linkage using a missing data formulation," *Adv. Stat. Anal.*, 2009, 93, 49-60.

Ghosh, S. K. and Norris, J. L., "Bayesian capture-recapture analysis and model selection allowing for heterogeneity and behavioral effects," *J. Agric. Biol. Environ. Stat.*, 2005, 10, 35–49.

Geneva Declaration, *Global Burdan of Armed Violence*, 2008, Geneva.

Gohdes, A., "Different convenience samples, different stories: The case of Sierra Leone," The Human Rights Data Analysis Group at Benetech 2010, http://www.hrdag.org/resources/publications/Gohdes_Convenience%20Samples.pdf

Harrison, M. J., O'Hare, O. E., Campbell, H., Adamson, A., and McNeillage, J., "Prevalence of autistic spectrum disorders in Lothian, Scotland: an estimate using the "capture-recapture" technique," *Arch. Dis. Child*, 2006, 91, 16-19.

Hook, E. B. and Regal, R. R., "Capture-recapture methods in epidemiology: Methods and limitations," *Epidemiologic Reviews*, 1995, 17, 243–264.

Hook, E. B. and Regal, R. R., "Validity of methods for model selection, weighting for model selection, and small sample adjustment in capture-recapture estimation," *Am. J. Epidemiol.*, 1997, 145, 1138–1144.

Hope, V. D., Hickman, M., and Tilling, K., "Capturing crack cocaine use: estimating the prevalence of crack cocaine use in London using capture-recapture with covariates," *Addiction*, 2005, 100, 1701-1708.

International Working Group for Disease Monitoring and Forecasting, "Capture-recapture and multiple record systems estimation I: History and theoretical development," *Am. J. Epidemiol.*, 1995a, 142, 1047–1058.

International Working Group for Disease Monitoring and Forecasting, "Capture-recapture and multiple record systems estimation II: Applications in human diseases," *Am. J. Epidemiol.*, 1995b, 142, 1059–1068.

Landman, T., *Studying Human Rights*, 2006, Routledge, London.

Landman, T., and Gohdes, A., "Multiple Systems Estimation and the Case of Peru," Research Note prepared for a meeting on casualty recording and estimation in conflict and post-conflict situations, hosted by Carnegie Mellon University and the University of Pittsburgh, 23-25 October 2009.

Landman, T., and Gohdes, A., "Principals, agents, and atrocities: The case of Peru 1980-2000," In *Civilian Casualties and Strategic Peacebuilding,* Seybolt, T., Fischhoff, Aronson, J., Eds, 2011, Oxford University Press, Oxford.

Laska, E. M., Meisner, M., Wanderling, J., and Siegel, C., "Estimating population size and duplication rates when records cannot be linked," *Statistics in Medicine*, 2003, 22, 3403-3417.

Lee, A.J., Seber, G.A.F., Holden, J.K., and Huakau, J.T., "Capture-recapture, epidemiology, and list mismatches: Several lists," *Biometrics,* 2001, 57, 707-713.

Lee, S-M., Hwang, W-H., and Huang, L-H., "Bayes estimation of population size from capture-recapture models with time variation and behavior response," *Statistica Sinica*, 2003, 13, 477-494.

Lum, K., Price, M., Guberek, T., and Ball, P., "Measuring elusive populations with Bayesian model averaging for multiple systems estimation: A case study on lethal violations in Casanare, 1998-2007," *Statistics, Politics, and Policy,* 2010, Vol. 1 : Iss. 1, Article 2. Available at: http://www.bepress.com/spp/vol1/iss1/2
DOI: 10.2202/2151-7509.1005.

Madigan, D. and York, J.C., "Bayesian methods for estimation of the size of a closed population,*"* *Biometrika*, 1997, 84, 19-31.

Manrique-Vallier, D., Ball, P., Price, M., and Gohdes, A. "Multiple-recapture techniques for the estimation of fatal victims in armed conflicts," In *Civilian Casualties and Strategic Peacebuilding,* Seybolt, T., Fischhoff, Aronson, J., Eds, 2011, Oxford University Press, Oxford.

Manrique-Vallier, D. and Fienberg, S. E., “Population size estimation using individual level mixture models,” *Biometrical Journal*, 2008, 6, 1051-1063.

Murphy, J. (2009): “Estimating the world trade center tower population on September 11, 2001: A capture-recapture approach,” *American Journal of Public Health*, 99, 65–67.

Murphy, J., Brackbill, R. M., Thalji, L., Dola, M., Pulliam, P., and Walker, D. J., “Measuring and maximizing coverage in the World Trade Center health registry,” *Statistics in Medicine*, 2007, 26, 1688-1701.

Papoz, L., Balkau, B., and Lellouch, J., “Case counting in epidemiology: Limitations of methods based on multiple data sources,” *Int J Epid*, 1996, 25, 474-478.

Silva, R. and Ball, P., The Profile of Human Rights Violations in Timor-Leste, 1974-1999, A Report by the Benetech Human Rights Data Analysis Group to the Commission on Reception, Truth and Reconciliation of Timor-Liste, 2006.

Silva, R. and Ball, P., “The demography of conflict-related mortality in Timor-Leste (1974-1999): Reflections on empirical quantitative measurement of civilian killings, disappearances, and famine-related deaths,” Chapter 6 in *Statistical Methods for Human Rights*, Asher, J., Banks, D., and Scheuren, F. J., Eds., 2008, Springer, New York, NY.

Smith, P. J., “Bayesian methods for multiple capture-recapture surveys,” *Biometrics*, 1988, 44, 1177-1189.

Spagat, M., Lau, F., Jewell, B., and Jewell, N. P. Cross Validation of Three Methods for Measuring War Deaths, 2011, to appear.

Spiegel, P. B. and Salama, P., “War and mortality in Kosovo, 1998-99: an epidemiological testimony,” *The Lancet*, 2000, 355, 2204-2209.

Tancredi, A., “Bayesian approaches to matching and size population problems: A unified framework,” paper presented at conference at the Politecnico di Milano: www2.mate.polimi.it/convegni/viewpaper.php?id=155&print=1&cf=7

Tancredi, A., “Capture-recapture models with matching uncertainty,” paper presented at the 2010 meeting of the Italian Statistical Society (SIS) in Padua: homes.stat.unipd.it/mgri/SIS2010/Program/18-SSXVIII_Luzi/817-1504-1-DR.pdf.

Tilling, K., “Capture-recapture methods—useful or misleading,” *Int J Epid*, 2001, 30, 12-14.

Williams, M., “Artificial Societies and Virtual Violence,” *Technology Review*, 2007, 110(4), 74-76. Online at http://www.technologyreview.com/Infotech/18880/.

York, J. C. and Madigan. D., “Bayesian methods for estimating the size of a closed population,” Technical Report 234, 1992, University of Washington.

Zwane, E and van der Heijden, P. G. M., “Population estimation using the multiple system estimator in the presence of continuous covariates”, *Statistical Modelling*, 2005, 5, 39-52.

Zwierzchowski, J. and Tabeau, E. M., “Census-based multiple system estimation as an unbiased method of estimation of casualties’ undercount”, Conference paper for the European Population Conference, 2010, Online at http://epc2010.princeton.edu/download.aspx?submissionId=100880.