Published in final edited form as:
Annu Rev Stat Appl. 2018 March ; 5: 95–118. doi:10.1146/annurev-statistics-031017-100641.

Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy

Sheila M. Bird[1,2] and **Ruth King**[3]

[1]MRC Biostatistics Unit, University of Cambridge School of Clinical Medicine, Institute for Public Health Cambridge CB2 0SR

[2]University of Edinburgh, Usher Institute of Population Health Sciences and Informatics, Edinburgh EH16 4UX

[3]University of Edinburgh, School of Mathematics, Edinburgh EH9 3FD

Abstract

Estimating population sizes has long been of interest, from the estimation of the human or ecological population size within regions or countries to the hidden number of civilian casualties in a war. Total enumeration of the population, for example, via a census, is often infeasible or simply impractical. However, a series of partial enumerations or observations of the population is often possible. This has led to the ideas of capture-recapture methods, which have been extensively used within ecology to estimate the size of wildlife populations, with an associated measure of uncertainty, and are most effectively applied when there are multiple capture occasions. Capture-recapture ideology can be more widely applied to multiple data-sources, by the linkage of individuals across the multiple lists. This is often referred to as Multiple Systems Estimation (MSE). The MSE approach has been preferred when estimating "capture-shy" or hard-to-reach populations, including those caught up in the criminal justice system; or homeless; or trafficked; or civilian casualties of war.

Motivated by a range of public policy applications of MSE, each briefly introduced, we discuss practical problems with potentially substantial methodological implications. They include: "period" definition; "case" definition; when an observed count is not a true count of the population of interest but an upper bound due to mismatched definitions; exact or probabilistic matching of "cases" across different lists; demographic or other information about the "case" which may influence capture-propensities; required permissions to access extant-lists; list-creation by research-teams or interested parties; referrals (if presence on list A results - almost surely - in presence on list B); different mathematical models leading to widely different estimated population sizes; uncertainty in estimation; computational efficiency; external validation; hypothesis-generation; and additional independent external information. Returning to our motivational applications, we focus on whether the uncertainty which qualified their estimates was sufficiently narrow to orient public policy; and, if not, what options were available and/or taken to reduce the uncertainty or to seek external validation. We also consider whether MSE was hypothesis-generating: in the sense of having spawned new lines of inquiry.

sheila.bird@mrc-bsu.cam.ac.uk; ruth.king@ed.ac.uk.

Keywords

confidentiality; deductive disclosure; demographic factors; evidence-based policy; hidden populations; quantifying uncertainty; record-linkage

1 Brief history of Multiple Systems (or Capture-Recapture) Estimation

We briefly outline the history of multiple systems (or capture-recapture) estimation of population sizes. The approach has been applied in numerous areas, from wildlife populations (King and Brooks, 2008) to casualties in war (Ball *et al*, 2003) and the number of pages on the World Wide Web on a given topic (Fienberg *et al*, 1999). The underlying concept for the simplest dual system estimation (where there are two partial enumerations of the population) is intuitive enough to be used in public outreach science events. For example, a one-minute game with plastic ducks was designed for the Cambridge Science Festival to show even young children how counting the overlap between their two independent duck-captures allowed estimation of the total number of ducks in a closed population, that is: the bucket the ducks were selected from. See also https://www.youtube.com/watch?v=aiSKgIc_8vk.

The idea of combining information from two different partial enumerations has a long history - and dates back to at least Graunt in the 1600s who applied the basic idea to estimate the effect of the plague on the population of England (Hald, 1990). However, perhaps the most famous early application of this dual system estimation approach came nearly 200 years later when Laplace used the approach to estimate the total population of France in 1802 (Manly *et al*, 2005). In this instance, two partial captures were used corresponding to (i) birth records of babies born across the whole of France; and (ii) census counts for several municipalities in France where local mayors conducted a complete census. Cross-classifying individuals recorded on the two surveys led to the following data:

- total of 1 million (approximately) individuals recorded on the national birth certificates;
- total of 2,037,615 recorded in the census of the given municipalities;
- 71,866 individuals recorded on both the birth certificates and census records.

Equivalently we can record these data as:

- 928,134 (= 1 million – 71,866) individuals recorded on the national birth certificates but not the census records;
- 1,965,749 (= 2,037,615 – 71,866) individuals recorded in the census of the given municipalities but not the birth certificates;
- 71,866 individuals recorded on both the birth certificates and census records.

Thus, also giving a total of 2,965,749 unique individuals observed.

The data are most easily presented in an incomplete 2 × 2 contingency table corresponding to the number of individuals observed by each distinct combination of surveys (or sources):

		Municipalities	
		0	1
Birth	0	?	1,965,749
Records	1	928,134	71,866

In the table the level of “0” corresponds to not being observed by the given survey; and a “1” for being observed. However, the number of individuals not observed by either survey is unknown (i.e. cell entry (0, 0)). We can estimate the total population size using the following argument. Consider (i) the proportion of individuals recorded by the municipality censuses that are also recorded on the birth certificates; and (ii) the proportion of the total population that are also recorded on the birth certificates. Assuming that the two partial enumeration processes are independent of each other we would expect these observed proportions to be approximately equal. Thus, equating these proportions and rearranging the expression provides an estimate of the total population.

Mathematically, let $n_{1,}$ and $n_{.1}$ denote the total number of individuals observed by survey 1 and 2, respectively; and let n_{11} denote the number of individuals observed by both surveys 1 and 2. Finally let N denote the total number of individuals. Applying the above rationale we obtain an estimate of N, denoted $\hat{N}$ by:

$$\widehat{N} = \frac{n_{1.}n_{.1}}{n_{11}}.$$

Applying this approach to the data collected by Laplace we have $n_{1.} = 1, 000, 000$; $n_{.1} = 2, 037, 615$; and $n_{11} = 71, 866$, leading to an estimated population of France of 28.35 million. Using a (modern day computationally intensive) non-parametric bootstrap algorithm provides an associated 95% confidence interval of (28.16 million, 28.55 million). The estimate obtained by Laplace using the dual estimation approach is similar to other published population estimates around that time (e.g. 27.5 million in 1801; Grigg 1980).

The above estimate by Laplace is an early example of what is typically referred to as the Lincoln-Petersen estimator, which was developed for estimating population sizes within fisheries (Lincoln, 1930; Petersen, 1896). For a history of this estimator, see Goudie and Goudie (2007), who also make the observation that Petersen was not the first to apply this approach. The Lincoln-Petersen estimator is a consistent estimator of the total population size, but biased for small sample sizes. This led to the Chapman estimator (Chapman, 1951) that corrects for the bias, providing the less biased estimator for the total population size:

$$\widehat{N} = \frac{\left(n_{1.} + 1\right)\left(n_{.1} + 1\right)}{\left(n_{11} + 1\right)}.$$

However, without additional information, the assumption of independence between the two surveys cannot be removed, or even tested, and so these estimators are limited in their applicability. The two survey approach was first extended to allow for K independent samples where the number of individuals sampled at each survey is fixed in advance of the study (Schnabel, 1938). This approach is often referred to as a Schnabel census. However, in general, although the number of surveys is typically fixed in advance, the number of individuals sampled is random for each survey, leading to further mathematical developments and a general multiple survey approach (Darroch, 1958). This approach permitted distinct capture probabilities for each survey, and an expression for the maximum likelihood estimate. The corresponding data for the multiple survey approach are the capture histories of each individual observed within the study, detailing whether or not the individual was observed by each survey. The data are usefully summarised in the form of an incomplete contingency table corresponding to the number of individuals observed by each distinct combination of surveys.

Due to the differences in the data structures, there was something of a divergence in mathematical developments between the ecological and epidemiological applications in the 1960s and 1970s. For ecological capture-recapture studies, the data collection is typically a temporal process whereby the surveys correspond to a series of discrete capture occasions over a given period of time. The first time an individual is observed, a mark is applied (such as a tag/ring) that can be uniquely identified at subsequent capture occasions; more recently, photographic identification may be used rather than a mark applied. Within epidemiological studies, the surveys typically correspond to a set of different lists or sources that are collated. Individuals are identified via unique identifiers (such as name, date of birth, address). An individual is simply recorded by a given source if they are observed within a specified time period; thus, the temporal information of the surveys is typically discarded. The presence/absence of the temporal aspect of the surveys is perhaps the most distinguishing difference between ecological and epidemiological capture-recapture studies/models. For the ecological applications, due to the temporal aspect of the studies, often over a longer period of time, the assumption of closure (no births/deaths/migration) was relaxed leading to the Cormack-Jolly-Seber (CJS) model (Cormack, 1964; Jolly, 1965; Seber, 1965); see King (2014) for a review of such open population models. For closed population models, in addition to survey dependence (i.e. time dependence), the capture probabilities were extended to permit trap dependence (removing the independence assumption from the capture probabilities, permitting a ‘trap happy’ or ‘trap shy’ response following initial capture i.e. behavioural effects) and individual heterogeneity. These dependencies are usefully summarised and described by Otis *et al* (1978). For a detailed discussion of these models, and associated extensions, see for example King (2014); McCrea and Morgan (2014).

In parallel, models were developed for the epidemiological framework to allow for additional individual heterogeneity via latent class models (Goodman, 1974), whereby the population is assumed to be composed of a set of sub-populations (strata), such that each sub-population (stratum) is homogeneous and within each stratum the associated surveys are independent of each other. Alternatively, the seminal paper by Fienberg (1972) introduced the concept of log-linear models, specifying the expected cell counts to be of log-linear form, allowing for interactions between the surveys, and hence removing the previous independence assumption. These log-linear models are the basis of the majority of multiple systems estimation applied to epidemiological data. For example, consider the case where there are $K = 3$ surveys, so that there are seven observable combinations of sources that an individual may be observed by. We let the set of such combinations be denoted by $R = \{100, 010, 110, 001, 101, 011, 011\}$, where the *i*th digit of each triple corresponds to being observed (=1) or not observed (=0) by survey *i*. We let n_{ijk} denote the number of individual observed by the combination of sources denoted by $ijk \in R \cup \{000\}$. We can again represent the data in the incomplete contingency table as in Table 1.

We assume that the cell counts are independent, conditional on the model parameters, such that,

$$n_{ijk} \big| \mu_{ijk} \sim Poisson\left(\mu_{ijk}\right)$$

where the expected cell means are of log-linear form:

$$\log \mu_{ijk} = \theta + \theta_i^1 + \theta_j^2 + \theta_k^3 + \theta_{ij}^{12} + \theta_{ik}^{13} + \theta_{jk}^{23}.$$

The term θ represents overall abundance; the main effect terms $\left(\theta_i^1, \theta_j^2, \theta_k^3\right)$ reflect the propensity of being observed by each source; and the two-way interactions $\left(\theta_{ij}^{12}, \theta_{ik}^{13}, \theta_{jk}^{23}\right)$ the dependence between each survey pair. This is the saturated model as, for the incomplete table with K surveys, it is not possible to estimate the K-way interaction term. Constraints are specified on the parameters to provide a unique representation (and unique maximise likelihood estimates). Sub-models are obtained by setting log-linear terms equal to zero. The estimate of the total population size is generally dependent on the fitted log-linear model, in terms of interactions present. Identifying the interactions present between the different surveys (and their direction i.e. positive/negative) can also be of interest for public policy to understand how individuals in the population interact with the different surveys. Sandland and Cormack (1984) showed that the alternative multinomial model specification is equivalent to the Poisson model specification, when conditioning on the total number of individuals observed. Log-linear models have been applied in numerous contexts and extended/adapted to allow for additional factors. For example, Tilling and Sterne (1999) discuss the inclusion of continuous covariates to allow for individual heterogeneity using a multinomial logit model. Alternatively, for discrete covariates, such as gender or age-group, the incomplete contingency tables may be stratified. Each stratified table may be analysed independently, although this can lead to relatively few overlaps between sources so that the statistical techniques cannot be sensibly applied; ignores common information or structure; and makes for additional complexity in correctly obtaining associated uncertainty intervals on the total population size (i.e. the sum of the individuals across all stratified tables). King *et al* (2005) directly address this issue by analysing all strata simultaneously, permitting the borrowing of information across the strata by specifying an extended log-linear model, treating the additional discrete covariate information as additional factors in the contingency table, leading to multiple unknown cells. The log-linear models are specified to allow additional interactions: survey × factor and factor × factor interactions. In addition, Bayesian approaches have been developed, permitting both the inclusion of prior information on the total population size and/or log-linear terms; and Bayesian model-averaging, incorporating both parameter and model uncertainty (Fienberg *et al*, 1999; King and Brooks, 2001a,b; King *et al*, 2005; King, 2014; Knuiman and Speed, 1988; Madigan and York, 1997; Overstall and King, 2014a,b).

2 Applications of Multiple Systems Estimation

We introduce six applications of MSE which have policy implications ranging from criminal justice, environmental and financial to human rights, public health and socio-economic. In these applications, the owners of data-sources (or lists) to be linked may cleave to different operating principles (harm reduction, risk-aversion, retribution, value-for-money) that need to be bridged for the common good of MSE to proceed. And MSE may itself be intermediary, by providing denominators (say), en route to additional calculations, of death-rates (say) for which numerators - as external data - were already known.

2.1 Statistical Ecology

Estimating population sizes can be vital for conservation and/or management. Population size is one of the factors used in classifying species on the International Union of the Conservation of Nature Red List (http://www.iucnredlist.org), in addition to others such as rate of decline and geographical distribution. Accurate estimation of abundance can be difficult for hard-to-find species, leading to greater uncertainty about classification and comparative ranking of endangered or threatened species. This often necessitates a range of different data collection techniques being used, for example: capture-recapture surveys, distance sampling, aerial surveys or combining different forms of surveys. Changing climate and habitat loss/fragmentation pose a particular threat to many species. For example, of the 17 species of gibbons, 11 are listed as endangered; four as critically endangered; one as vulnerable and one with data too deficient to classify. All the classified gibbon species had a decreasing population trend, primarily due to habitat loss and hunting. Understanding the whole ecosystem, and the factors driving such populations can be important for conservation/management purposes, in order to predict effects related to, for example, changing food availability and/or habitat (say).

2.2 Persons who inject heroin

Heroin injectors incur criminal justice, welfare and healthcare costs from the public purse (White *et al*, 2014); have reduced quality of life; and experience a high rate of premature mortality - especially soon after prison-release or hospital-discharge, and with variation by gender and age-group (Bird and Hutchinson, 2003; King *et al*, 2013, 2014; Merrall *et al*, 2010, 2012; Pierce *et al*, 2015; Seamean *et al*, 1998; White *et al*, 2015). Injectors are difficult to count for both because they exist on the fringe of legality and because the intensity of their heroin injecting varies over time. For example, intensity of injecting is markedly reduced if an individual is currently receiving opioid substitution therapy (OST); or in prison; or hospitalized on account of a non-fatal overdose, mental ill-health, external injury, blood-borne viruses, other infectious diseases, or for respiratory and liver diseases, which commonly occur in such populations.

Different lists pertaining to (heroin) injectors are available in Scotland and England (King *et al*, 2009a, 2014). One of Scotland's lists was its confidential register of Hepatitis C virus (HCV) diagnoses. Over the years, the proportion of new HCV diagnoses with undeclared risk-behaviour has decreased and Scotland's HCV Action Plans have successfully promoted confidential HCV testing of persons born in 1956-75 who had ever injected, as a high proportion were expected to be HCV antibody positive (Hutchinson *et al*, 2005). However, this has led to further issues as, increasingly, those who declared injecting as their HCV risk-behaviour were former, not current, injectors (see Section 3.4 for further discussion).

2.3 Problem drug users

Case definitions of problem drug users (PDUs) differ: between nations, over time within a nation, or across the lists currently used to estimate a national count of PDUs. For example, recently but not historically, Scotland defined its PDUs as regular users of illegal opioids or benzodiazepines; or patients prescribed methadone in the treatment of their addiction (thereby including persons in receipt of OST); whereas England's definition was users of illegal opioids or crack-cocaine. Representatively-sampled, household-based surveys of 16-59 year olds, such as the Crime Survey for England and Wales (2016) which includes computerized questions on past-year use of illegal drugs, have been considered as alternative to MSE for estimation of PDUs. However, household-based surveys seriously under-estimate past-year use of hard drugs such as heroin because a notable proportion of users are homeless or incarcerated.

In England and Wales, there is mandatory salivary testing (for opiates and/or cocaine) of those arrested for a list of trigger offences. The policy's intention was referral of those testing positive to drug treatment agencies. However, Jones *et al* (2014) have shown that the more successfully the referrals (or "scooping" up) of individuals who test positive is at engaging them in drug treatment, the more problematic for MSE as these individuals who test positive are, almost surely, also listed as drug treatment attenders. Individuals who test positive and individuals on treatment lists are thereby overly inter-dependent (in one direction at least).

2.4 Homeless

Censuses of homeless individuals in a defined district and period (midnight to 3am on the census-day, say) will typically lead to an undercount of the true population. For example, volunteers that enumerate the number of individuals within a given region and time period at hostels or on the streets will typically miss homeless persons who are temporarily accommodated (in hospital or custody, say). Further, the volunteers cannot ascertain rough-sleepers' age-group, or other demographic data, without awakening them. As an alternative Fisher *et al* (1994) used six source-lists and applied an MSE approach to estimate the number of homeless individuals in north east Westminster.

Plant capture approaches have also been applied in the USA, which rely upon a single capture occasion in a given area and time (Laska and Meisner, 1993). In such a study, volunteers are “planted” into the community and report whether they have been observed during the single survey, leading to application of a dual systems approach in which the planted volunteers are treated as “marked” individuals prior to the survey.

2.5 Human trafficking

Victims of human trafficking have generally been duped, incentivised, captured, coerced or brutalized into leaving their homeland. Their exploitation - typically for prostitution, drug trafficking or domestic slavery - is secured by the impounding of their identity papers, impoverishment to repay alleged debts, violence (and the fear of being killed) and often a language barrier, see https://www.theguardian.com/uk/2011/feb/06/sex-traffick-romania-britain.

Whether, in the wider public interest, the victims of human trafficking may be excused prosecution for serious crimes ranging from drug trafficking up to manslaughter that they have been forced, or driven, to commit is uncertain, see http://www.carmelitechambers.co.uk/news-and-events/news/human-trafficking-victims-should-not-be-charged-with-murder-felicity-gerry; and further inhibits victims from seeking to escape those who control them.

Victims of human trafficking are hidden because those who exploit them want to remain below the radar of police, customs, hospitals, and landlords. Nonetheless, Silverman (2014) successfully applied an MSE approach using five lists which enumerated 2,744 potential victims of trafficking into the UK and estimated a further 7,000 to 10,000 as the unenumerated dark figure.

2.6 Crimes against humanity

The work of the Human Rights Data Analysis Group (HRDAG), now in existence for a quarter century (“25 years and counting”), has featured in high-profile legal cases, such as the conviction in Guatemala of General ER Montt for genocide and crimes against humanity, and has underpinned Truth and Reconciliation Commissions. In Peru, for example, HRDAG estimated that around 70,000 deaths had occurred, of whom only 25,000 had been documented directly (18,000 by the Commission, another 7,000 by data-sources other than the Commission). See Seybolt *et al* (2003) for the steps that HRDAG takes to ensure that its MSE methodology and machine-learning algorithms for reproducible matching can withstand cross-examination.

Briefly, in the human rights context, each data-source for MSE is expected to record the names of those whom it lists as having died. Datasets are first checked for clerical and logical errors, and duplicates are removed. As the same information (for example, on sex) may be coded differently across datasets (male/female, m/f or 1 vs 2), the next step is to synchronize coding across relevant datasets. Alignment may lead to some degradation if the datasets have recorded information differently (child/adult or <15 years, 15-44 years, 45-64 years, 65+ years). If there has been no pre-agreement across diverse data-sources on the terms to be used in describing human rights violations, the MSE analyst has to define a mapping from the terms adopted by each data-source onto a common vocabulary. For example, murder, homicide, lethal force map to “homicide”, say. Better by far if a common vocabulary can be agreed in advance; and used by each data-source.

Stratification of datasets, prior to identifying overlaps, is often prudent because the propensity for a homicide to be listed by specific data-sources may depend on sex and age-group as well as on geographical area in virtue of time-varying regional intensities of conflict, as in Syria or Afghanistan (Bird and Fairweather, 2009). Over-stratification can, of course, result in too few overlaps for MSE to be sensibly applied.

3 Recent methodological developments

3.1 Statistical ecology

Population estimates from capture-recapture studies can be highly dependent on the dependence of the capture probabilities of individuals, and the factors that may affect them. Incorporating individual heterogeneity has been of particular interest in order to reflect biological realism (individuals typically differ in their capture propensity). A number of different models have been developed in order to account for such heterogeneity. Pledger (2000) considers models akin to Goodman (1974), incorporating heterogeneity via the form of latent classes (or discrete mixture models), in addition to allowing for temporal and behavioural effects. Infinite mixture models are attractive due to their interpretability but have additional model-fitting complexities as, in general, the likelihood is analytically intractable, expressible only as an integral. Model-fitting tools for addressing this issue include the use of numerical integration (Coull and Agresti, 1999; Gimenez and Choquet, 2010) and Bayesian data augmentation (Durban and Elston, 2005; King and Brooks, 2008; King *et al*, 2009b); with associated efficient model fitting techniques developed (King *et al*, 2016). For a review of such models, see for example King (2014).

Advances in technology have led to new issues and associated statistical tools. For example, the identification of individuals using DNA matching from hair or scat samples, or photographic recognition and hence potential mismatching (Wright *et al*, 2009). Alternatively, acoustic recordings may be used to identify individuals, rather than physical re-sightings and identifiable characteristics. For difficult to observe populations, an array of motion sensor camera/acoustic traps may be erected that captures animals passing by. This generates further statistical issues with regard to spatial information leading to the development of spatially explicit capture-recapture modelling (Borchers and Efford, 2008; Efford, 2004; Royle *et al*, 2014) and continuous time observations (Borchers *et al*, 2014). For a review, see for example Borchers and Fewster (2016).

3.2 “Period” definition

The definition of period for which prevalence-estimation is required typically balances the time required for there to be a reasonable chance of “listing” the persons of interest by the various MSE data-sources and, on the other, timeliness for the prevalence-estimation to be policy-relevant. Thus, Scotland estimated its number of current injectors, by sex and age-group, at roughly three-year intervals because public health professionals were interested in knowing whether younger individuals were being dissuaded from injecting; and the extent to which older injectors were (or were not) aging out of injecting, or dying from overdose.

Of further interest is MSE of how many individuals who were "listed" as injectors in (say) 2008/9 are persistent in the sense of being also "listed" in 2011/12. Matching of "cases" across, not just within, periods is more tricky, requiring access to age in years, rather than just age-group, for optimal matching (exact or probabilistic) across periods. However, the research-team who compiles and assesses the overlaps between lists in 2008/09 may, of course, be different from the team funded to do so in 2011/12. Analysts typically receive only demographically pre-specified overlap-counts: for example, for each combination of region, sex and age-group. Persistence can only be addressed by revisiting the field-work for both 2008/09 and 2011/12; and only if both field-work teams have access to individual years of age to aid matching across periods, provided that clients did not also migrate much from one region to another. Migration to live away from fellow injectors is, however, a means to reduce clients' relapse into injecting.

3.3 "Case" definition

Scotland and England adopted different contemporary definitions for problem drug users (PDUs). Moreover, Scotland's "case" definition for PDUs has evolved over time as the inclusion of persons in receipt of opioid substitution therapy (OST) assumed greater importance than it had in the 20th century when OST recipients were proportionately fewer (Strang *et al*, 2010). By the 21st century, optimism that OST meant cessation from injecting had given way to greater realism: that OST reduces clients' frequency of injecting but they do not necessarily cease injecting.

"Case" definition may also be less well adhered to by some data-sources than others: for example, the risk behaviour that Scotland's confidential HCV register records is "ever-injector" (not "current-injector"). The distinction was less problematic in the early days of the HCV register when confidential HCV testing was mostly offered to current injectors but, in the past decade, has been targeted at older, former injectors whose HCV-related liver progression would be likely to need antiviral treatment.

A different problem arises when clients elect not to declare a risk-behaviour, such as injecting, which might otherwise explain their HCV infection. Hutchinson (2004) used an MSE approach to deduce that the vast majority of Scotland's undeclared-risk HCV diagnoses were injection-related. Thereafter, a decision has to be made as to whether "case" definition for injectors among Scotland's registered HCV diagnoses should include those whose risk-behaviour is undeclared.

3.4 Observed count as upper bound for count of interest due to mismatched definitions

When using Scotland's HCV database for estimating current injectors, Overstall *et al* (2014) had to account for the increasing number of recorded individuals that declared injecting as their HCV risk-behaviour but were former, not current, injectors. Existing MSE methodology was extended to account for the "injection-related HCV diagnoses not otherwise listed" being an upper bound for "current injectors' injection-related HCV diagnoses not otherwise listed". Such individuals, if they were recorded on any other of

Scotland's MSE lists for current injectors, were classed as current injectors. However, individuals who were not observed by any other MSE source were treated as a mixture of current and former injectors so that their recorded number is essentially an upper bound for the number of current injectors, instead of being their observed number. Overstall *et al* (2014) explicitly modelled this observed cell entry as a mixture of current and former injectors, applying a Bayesian data augmentation technique to impute the true number of current injectors within the associated Markov chain Monte Carlo algorithm. This analysis clearly demonstrated that failing to account for the additional complexity led to significant over-estimation of the total number of current injectors (by a factor of two for 2009) and a potential mis-interpretation of there being a constant population over time. The new methodology produced a progressive reduction in Scotland's estimated number of "current injectors" in concert with Scotland's Hepatitis C Virus Action Plans successful outreach to older former injectors by offering them confidential HCV testing.

3.5 Exact or probabilistic matching of "cases" across different lists

Amongst the best-documented examples of robust specification for how matching of "cases" across different lists will be judged are when MSE is used to quantify crimes against humanity, as the intention is to bring to justice the perpetrators of those crimes. Rigorous specification is essential, not least because the court's judgment on capital crimes may rest upon the rigour deployed.

In record-linkage studies, exact matching of "cases" across lists is generally considered less accurate than probabilistic matching because exact matching does not allow for inevitable (and often obvious) data-errors. Of course, probabilistic matching needs to be programmed so that the allowable extent of discrepancy for component i of the matching-string between the members of lists A and B is defined and the weighting of discrepancies between different components (say i and j) when comparing members $A(n)$ and $B(m)$ of lists A and B is also determined. Rigour is essentially codified by how the matching program is written. See, for example, Lee (2002) for discussion of the impact of matching errors and an approach to account for possible matching errors.

Statistical methods may envisage that the analysis-team has full access to all lists, A to D say, so that, according to some optimization criterion, they can update, and indeed optimize, the "match-weights" in-context as the estimation-task and list-propensities unfold (Harron *et al*, 2016). While this approach is academically attractive, it is not generally practicable and, to the extent that it is successful, may enhance the risk of deductive disclosure by the very optimality of its matching. Sutherland and Schwarz (2005) extend the standard framework where only partial matchings are available between lists for example where lists A and B are matched; and lists B and C are matched; but A and C are not able to be directly matched. The ideas can be extended further if lists are stratified, yet not all lists are active in all strata.

3.6 Demographic or other information about "case" influences capture-propensities

When list-holders and analyst-team differ, the analysts have to specify in advance the cross-classified covariate-strata, for example defined by sex (2 levels) × age-group (3 levels) × geography (3 levels) and so 18 strata, for each of which the 15 overlap counts across lists A

to *D* (say), need to be worked out and provided to the analyst-team for demographic capture-propensities to be investigated thoroughly.

Difficulties can arise for the analysts if the data-providers do not allow counts below five to be published on the grounds that press, police or others (notably holders of lists *A* to *D*) may be able to deduce previously unknown-to-them attributes of their list-members. For example, if all four young females in a particular geographical region on list *A* (methadone clients) were listed also as present on lists *B* (benefits recipient), *C* (child in care) and *D* (incarcerated in the past year), then the drug-treatment team for the geographical region in question could deduce that all four young women had a child in care, were in receipt of benefits and had been incarcerated in the past-year.

For the greater public good of making MSEs, analysts may have to accept limitations on the extent to which the cross-classified counts they received can be disclosed. In some cases, the complete data may be made available to researchers under confidentiality agreements, yet the data may only be published with the censored cell values. See King *et al* (2014) for MSE of injectors in England, where cells with observed counts of 1-4 were simply represented by * and, consequently, their results are not reproducible by others.

3.7 Permission to access extant-lists

Privacy access committees (PACs) have an important role in adjudicating the public good that a particular MSE represents and make their judgment on the basis of the public interest case that analysts or policy-makers advance. Approval by a competent PAC is necessary, but not sufficient, to guarantee access to the confidential lists cited by analysts as a sound basis for MSE. List-holders, *A* to *D* say, each need to agree to prepare their client-list suitably for transfer to a safe-haven for matching across the prepared lists to be programmed according to the PAC-approved rules which were defined by the analyst-team.

3.8 List-creation by research-team or interested parties

Hay *et al* (2009) made considerable efforts to assist local drug treatment teams in creating client-lists which were suitable for use in separate MSEs of Scotland’s number of injectors and of PDUs (appropriately defined). Subsequently, the Scottish government sought to utilize only lists that were held centrally and electronically, thereby saving costs. Estimation of current injectors was suspended, however, initially due to convergence difficulties but also in recognition that OST-clients, now counted as PDUs, might simultaneously qualify as currently injecting (Information Services Division, Scotland, 2016).

When interested parties, rather than an independent research-team, apply “case” definitions subconscious (or deliberate) bias may influence the listing of “cases”. Objectivity in the application of case-definitions is crucial but desperately difficult when MSE is used to quantify crimes against humanity and methods will ultimately be tested in court. For this reason, some of the clearest thinking on “case” definition and on algorithms for the matching of “cases” across available data-sources can be found in the MSE literature on human rights violations.

3.9 Referrals (presence on list *A*, almost surely, results in presence on list *B*)

Interested parties may contrive referrals between lists in a misguided, but generally detectable, effort at corroboration. Referrals can arise as a consequence of policy-decisions such as England's arrest-referral policy whereby those arrested for trigger-offences who tested positive for opiates or cocaine were to be referred to drug treatment teams which created a structural overlap between individuals who test positive for opiate/cocaine and drug treatment clients. Jones *et al* (2014) show that substantial bias can be introduced if standard models are applied when referral mechanisms exist between the data-sources.

3.10 Different model-frameworks giving rise to widely different estimations

King *et al* (2013, 2014) found that limitation on the allowable interactions (1st order; 2nd order; 3rd order) was more important in MSE than whether the analysis was conducted from a Bayesian or frequentist stand-point. The Bayesian framework has the advantage of allowing analysts: (i) to incorporate independent external information in terms of the total population size, interactions present in the model and the sign of the interaction (King *et al*, 2005); and (ii) to permit the calculation of posterior probabilities for each model that sits within the set of allowable interactions within a robust framework that also leads to model-averaged estimates of the population size, incorporating both model and parameter uncertainty (King and Brooks, 2001a).

However, when bi/multi-modality is observed in the marginal posterior distribution for the total population size, further investigation is judicious in order to identify why this may occur (for example due to inclusion/exclusion of a particular interaction) and it may not be prudent to present a single model-averaged estimate. In such cases, a fuller description of the posterior distribution would be a better summary, including, for example, a plot of the posterior density for the population size, together with probabilities associated with the range of values for the population size and/or an investigation of the different models that give population estimates in the different modes (King *et al*, 2005; Overstall and King, 2014a). In addition, the Bayesian framework ensures that the sum of (say) regional estimates accords with the national total and also that credible intervals can be computed for quantities of interest besides the number of current injectors, for example: drug-related death-rates per 100 current injectors (King *et al*, 2014).

Tensions arise, however, when policy-makers want MSEs for smaller geographical regions than the data-sources were gleaned from; or seek to persuade the data-collection teams to base their work on such small geographies that not only do costs increase but appropriate allowance for interactions is impossible due to reduced overlaps between sources. Hence, our preference is always to work on a geographical (or other) scale that supports due diligence in the investigation of capture-propensities - even in the knowledge that, subsequently, regional estimates will be "localized" by an untested assumption of homogeneity across small areas.

3.11 External validation

Empirical discoveries, including by MSE, ideally need external validation - by deploying the same or similar methodology in different jurisdictions, or in the same jurisdiction but for a different time-period, or by adopting a different study-method to test directly the deductions that followed from MSE.

In short, discoveries or deductions made from MSE may need external validation to be ultimately persuasive. For example, estimation of how many were unlawfully killed by a corrupt regime or in a war-torn country may be corroborated by the finding of mass graves and forensic determination of how the victims died. In the UK, MSE-deduction about the age-related loss of female advantage in terms of drugs-related deaths per 100 injectors was reinforced when a similar interaction was shown to apply in England (King *et al*, 2014). Subsequently, the role that OST might have in explaining this strong sex by age-group interaction was investigated in a record-linkage study on the opioid-specificity of deaths for some 33,000 methadone-prescription clients in Scotland (Gao *et al*, 2016).

The impact on the risk of drugs-related death of referrals into drug treatment those arrested for trigger offences who had tested positive for opiates or cocaine was investigated by Pierce *et al* (2016) who found that the benefit of drug treatment was significantly reduced, but not negligible, for clients referred into drug treatment by the criminal justice system.

3.12 Implementation of MSE approaches

The computational task of implementing MSE increases as the number of sources and covariates (or covariate levels) increases and additional model complexity is introduced, for example, to allow for extra unobserved heterogeneity or censored cells. To aid in the analysis of such data, a range of computer packages has been developed. See, for example, McCrea and Morgan (2014; Section 2.13) for a brief summary of many of these, particularly in relation to the ecological literature, where each package is typically targeted at a small set of specific types of model. For log-linear models, the `R` package `Rcapture` (Baillargeon and Rivest, 2007) fits a variety of models, calculating the associated MLEs within the classical framework, and a stepwise selection process can be used to determine the log-linear interactions present. Alternatively, within the Bayesian framework, efficient model-fitting algorithms have been of particular interest, to explore large model spaces in the presence of even a moderate number of surveys and covariates. The `R` package `conting` (Overstall and King, 2014b) provides posterior distributions for population sizes and interaction terms, and allows for multiple covariates with a general number of levels and for possible censored cells (as in Overstall *et al* (2014), with the observed cell count being equal to only an upper bound for the true value); and is implemented via the efficient reversible jump algorithm proposed by Forster *et al* (2012).

4 Validation of Multiple Systems Estimation: fit for public policy purpose?

Was the MSE uncertainty sufficiently narrow to orient public policy? If not, what options were taken to reduce uncertainty or seek external validation? As importantly, was MSE hypothesis generating in the sense of having spawned new lines of inquiry? We address these questions in further discussion of our motivating areas.

4.1 Statistical Ecology

Within ecological studies, multiple data-sources may be collected on the same population. For example, capture-recapture data; nest record data (such as number of eggs produced, or number of chicks fledged etc); population counts; dead recoveries. Many of these data-sources provide information on the same model parameters, such as population size, survival probabilities, fecundity rates etc. Analysing each dataset independently ignores the overlapping information available and can lead to post-hoc comparisons of estimates and/or combination of estimates. Integrated data analyses provide a robust mechanism for simultaneously combining the different forms of data within a single robust analysis. This permits the borrowing of information across the different data-sources, thereby increasing the precision (potentially significantly) of the parameter estimates (Besbeas *et al*, 2009; Brooks *et al*, 2004). In addition, comparing the integrated estimates with those obtained from each individual analysis may provide evidence of inconsistencies which can be investigated further, providing novel insight into the ecological system; and the model adapted accordingly (Reynolds *et al*, 2009). Finally, we note that in some cases not all of the capture occasions may be used in the statistical analysis: for example, due to particular survey protocols or to permit the assumption of closure. In such cases it may be possible to use the additional information to provide some validation of the population estimates (Worthington *et al*, 2017).

4.2 Persons who inject drugs

Scotland’s serial Bayesian MSE by sex and age-group (and region) of persons who inject drugs began in 1999/00 and ended in 2009/10 when estimates were produced for problem drug users (PDUs) but not for current injectors. For each such estimation, the MSE uncertainty was wide but not so wide that one could not discern: diffusion of Scotland’s injectors away from the central regions; different sex-distribution by age-group; and, in later estimations, lower number of injectors recruited into the youngest age-group as the next generation had been, to a notable extent, dissuaded from injecting. However, there were other concerns about MSE of Scotland’s injectors: first the presumption that most drugs-related deaths (DRDs) occurred in current injectors resulted in estimated DRD-rates per 100 current injectors that, even for Scotland, were very high. Subsequently, Scotland’s definition of current injectors has evolved to include methadone-clients, who nonetheless experience DRDs (Gao *et al*, 2016). However, cohort studies in Scotland and England of drug treatment clients who were prescribed opioid agonists have also confirmed that the DRD-rate (and methadone-specific death rate) for Scotlands clients is substantially higher than for their counterparts in England (Pierce *et al*, 2017).

Unsurprisingly perhaps, Scotland’s more recent MSEs have focused on problem drug users who include methadone clients, and have given up on estimation of current injectors. There were two reasons for giving up. First, the new methodology by Overstall *et al* (2014) to enable Scotland’s registered HCV diagnoses with injecting as their risk-factor to continue as a fourth data-source for injectors led to a substantial reduction in Scotland’s centrally-estimated number of current injectors. Secondly, the MSE attempted in 2009/10 using just three centrally-available electronic data-source either failed to converge or gave answers that the review-group was not confident about. MSE of current injectors has not been reported for 2012/13 (Information Services Division, Scotland, 2016).

As described in the previous section, the sex by age-group interaction in respect of Scotland's DRDs per 100 current injectors was corroborated in England and spawned further research to explain this validated empirical discovery. Meanwhile, Scotland's serial MSEs (with uncertainty) for current injectors by sex and age-group were used by Prevost *et al* (2015) in their multi-parameter evidence-synthesis to estimate Scotland's prevalence of Hepatitis C virus carriage in former injectors; and demonstrated sensitivity to their inclusion. The hoped-for refinement of MSEs by incorporating the age-specificity of HCV diagnoses was not forthcoming.

4.3 Problem drug users

The UK Advisory Council on the Misuse of Drugs (2000) had called for DRD-rates to be reported per 100 PDUs rather than per 100,000 of population as PDU-prevalence was not homogeneous across the UK.

Scotland's central MSEs (with uncertainty) for its PDUs have been in the range 55,000 to 60,000 for over a decade, albeit with some changes apparent in the demography of PDUs. Scotland's PDU estimations have been essential for policy-makers in: the resourcing of local drug action teams; the monitoring of methadone-prescriptions per 100 PDUs; and in the setting of community-based targets for Scotlands National Naloxone Programme so that, by the end of 2015/16, naloxone-kits had been supplied to 30% of a region's PDUs (Bird *et al*, 2017). The PDU estimates (with uncertainty) also feature in Scotlands official statistics on DRD-rates per 100 PDUs, which are reported by i) sex and age-group; and ii) NHS region (National Records of Scotland, 2016). By contrast, Millar and McAuley (2017), on behalf of the European Monitoring Centre for Drugs and Drug Addiction, are the first to attempt the reporting of opioid-deaths per 100 PDUs for different member-states of the European Union, a difficult but worthwhile task.

4.4 Homeless

In UK at least, central estimates (with uncertainty) for the number who are homeless in major cities have not hit the headlines. Nor has there been much discussion about other approaches: for example, 30% of 1947 current injectors interviewed in 2015/16 by Scotland's Needle Exchange Surveillance Initiative, which is geographically-representative, reported that they had been homeless in the past 6-months (Bird *et al*, 2017) but interviewees were not asked for how long they had been homeless (Health Protection Scotland, 2017). Surveys of prison inmates or benefit claimants might pose the same question and a patchwork of estimates, taking overlaps between respondents into account, might be stitched together. Charities or churches who assist the urban homeless might arrange an annual survey of locations where the homeless might spend the night so that trends in their number, sex and age-distribution, and possibly also in other risk-factors (such as intoxication) might be monitored.

4.5 Human trafficking

Silverman (2014), while chief scientific advisor at UK's Home Office, used MSE based on five data-sources to provide a preliminary, but shocking, first estimate of the extent of human slavery in the UK in the 21st century. As a consequence of there being between 10,000 and 13,000 potential victims of human trafficking in England and Wales in 2013, of whom fewer than 3,000 were listed, the next estimation for the Home Office may want to consider whether account can be taken of whether victims were identified singly or in a cluster; and of the sex, age-group (child versus adult) and other characteristics (continent of origin, say) of the listed potential victims of trafficking. As Silverman (2014) points out, consent is not required from child-victims for information about them to be reported to government agencies. Hence capture-propensities may be different for child versus adult victims.

By considering how many deaths due to external causes have occurred to potential victims of trafficking in a 5-year period and by eliciting expert opinion (say) on the plausible external-cause death-rate which might apply to potential victims of trafficking (the same as for injectors, perhaps), comparison could be made between different methods of estimation; or the death-rate estimate (with uncertainty) could be incorporated in Bayesian MSE as prior information on the total count of victims. Expert opinion on victims' death-rate from external causes might include suitably sensitive questioning of surviving victims about their peers.

4.6 Crimes against humanity

Using MSE and based on 4,400 documented deaths, Ball *et al* (2003) estimated in 2002 that approximately 10,000 Kosovar Albanian civilians (95% credible interval: 9000 to 12,000) were killed during March to June 1999, a claim disputed by the defence at the International Tribunal for the Former Yugoslavia but corroborated by a survey-based estimate, see Spiegel and Salama (2000), of 12,000 fatalities (95% confidence interval: 5,000 to 18,000), most during March to June 1999. By 2011, the Kosovo Memory Book had documented 14,000 deaths and disappearances, some of them military, between January 1998 and December 1999: again, most during March to June 1999.

5 Parallels and distinctions: Multiple Systems Estimation and Record-linkage studies

Fienberg *et al* (1999) noted how intimately entwined in practice were MSE, record linkage and missing data and therefore provided an integrated methodology for MSE and record linkage using a missing data formulation. Here, we focus on parallels and distinctions at the practical, rather than technical, level but recognise that well-understood practicality must ultimately have a technical translation for there to be realistic modelling of complexity.

5.1 Limited number of estimation goals

For human populations, MSE and record-linkage studies both typically require a study-protocol which clearly defines eligibility together with the identifying, demographic or other covariate information that will be used for "matching" across lists or databases; and the data-items from each list/database (for example, sex and birth-year; year of first live birth; year of first incarceration; year of starting to inject; year of HCV diagnosis) that may be used, inter alia, for quality-assurance that the linked-records, on a probabilistic basis, seem to pertain to the same individual. Thus, Merrall *et al* (2012) identified discrepancies, including in the reported year of starting to inject across serial episodes of drug treatment for clients of the Scottish Drug Misuse Database, but chose to make face-validity correction for sex only - in the event of pregnant males!

Unlike record-linkage studies which may have a range of estimation goals (dependent upon how strictly the granted-approval ensures that every requested data-item is accounted for in the analysis plan), MSE studies must specify quite precisely the covariate-strata, for each of which the research-team requires the overlap counts across its nominated lists. Alternatively, the MSE research-team may itself receive, from diverse sources, the source-list of (partially or wholly) identifiable individuals together with covariate information to enable the research-team to program its own matching across source-lists as the extent of matching may be dependent upon the quality of covariate information received: for example, male child, female child, adult male, adult female without the ability to make further differentiation by age. See also Sutherland *et al* (2007) for further discussion where lists may not be active for all strata.

5.2 Counts versus serial event-dates

A key distinction between MSE and record-linkage studies is that, whereas MSE is often done serially (for example, once every 3-years, say, for a single species or for several), record-linkage studies typically focus on an evolving time-sequence of events (different per linkable source-list, for example: drug treatment episodes, incarcerations, HCV diagnosis, live births, methadone prescriptions, cause-specific hospitalizations) that eligible clients have, or have not, experienced.

Methods of MSE which account for the 3-year persistence of persons who inject drugs could be devised but typically would require that the research-team has access to source-lists from 3-years previously and currently. Often, however, source-lists are administrative and, as such, the information on listed individuals is updated/corrected without there being any date-stamp on the changes made. Thus, MSE of 3-year persistence may require comparison between the archived source-lists from 3-years ago and its current format as well as the present-day source-list. On the other hand, record-linkage studies which focus on event-sequences may come to borrow ideas from MSE in terms of missed-recording of events.

5.3 Risk of deductive disclosure about individuals

Serial event-dates per linkable data-source give rise to immense concern over deductive disclosure about individuals - were the pseudonymised linked data (that is: across data-sources A, B, C, D) to be accessible by any of the contributing data-sources for the simple reason that an individual’s event-sequence on any of the data-sources (say A) may be sufficient for that data-source (A) to de-identify the person and thereby to learn about the A-client’s event-history in terms of B, C and D: which the de-identified individual may have wished to keep confidential from the list-A holder. To protect against deductive disclosure as described, the Farr Institute in the UK has established safe havens in which the linked-database can be analysed but from which copies of the linked database cannot be removed; nor, of course, made Open Access.

Deductive disclosure about individuals is also a risk in MSE studies when source-lists are centralised for the purpose of determining the cross-counts for analysis. Particularly when estimating how many persons have been victims of crimes against humanity, the very creation of bespoke-lists may endanger (by de-identification) their compiler.

5.4 Risks of redress

Record-linkage studies risk making discoveries about criminal offences committed by anonymised subjects, or "cases", who may, as a consequence, be liable to prosecution if a court-order has been obtained which requires their identification. Similar concerns may apply in MSE if the research-team holds sensitive lists of identifiable individuals.

Bird and colleagues curtailed a record-linkage study (Hutchinson *et al*, 1998) which had periodically matched the master-indices of 636 former at-risk Glenochil prisoners against Scotland's similarly indexed register of HIV diagnoses to discover if any were subsequently HIV-diagnosed. And if so, whether on the basis of their HIV diagnostic sample, they had been infected with the same HIV virus as had 13 of the 14 HIV seroconversions that were reported by an Infection Control Exercise at Glenochil Prison (Taylor *et al*, 1995). However, around 20 HIV infections may have occurred in Glenochil Prison during March to June 1993, more than were diagnosed at the time (Gore *et al*, 1995), as only two-fifths of inmates had agreed to have a personal HIV test during the Infection Control Exercise and prisoners who had been released or transferred from Glenochil Prison were not contacted.

Despite a good match, the research-team suspended its study in 2001 when a former Glenochil inmate, Stephen Kelly, was sentenced to 5 years' incarceration for having culpably and recklessly transmitted HIV infection to a female sexual consort (Bird and Leigh Brown, 2001). In Scots law, two independent pieces of information are needed for a conviction. The uniqueness of Mr Kelly's date of birth among the 636 former Glenochil inmates and the laboratory's retained date-of-birth label on the Glenochil HIV test samples were sufficient to identify Mr Kelly as one of the 13. His female sexual partner had also been infected by the "Glenochil virus".

Record-linkage research teams face sanctions in the UK if they deliberately seek to de-identify subjects whose anonymity they have undertaken to uphold. Deductive disclosure may occur inadvertently, however. One of 97 audited fatal accident inquiries into prisoner deaths in Scottish prison custody in the first five years of the 21st century (Bird, 2008) concerned a man who had become HIV-infected in Glenochil Prison during March to June 1993, was at liberty for only a few months, then re-incarcerated in Shotts Prison where he was diagnosed with AIDS lymphoma and died in 2001. Bird (2008) had inadvertently discovered the 14th man in the Glenochil cohort. He had never sought HIV treatment and, sadly, died eight and a half years after his HIV infection.

In bringing to justice perpetrators of crimes against humanity, MSE-investigators need to be scrupulous in their methodology from "case" definition through to matching programs so that their evidence holds up under cross-examination; and does not put in further jeopardy survivors of crimes against humanity. See https://hrdag.org/2013/03/11/mse-the-basics/ for some notable convictions.

5.5 Application to families, not individuals

Crimes against humanity and potential victims of human trafficking may be clustered, within families (say), so that MSE may occasionally need to consider family versus individual-members within a household as the unit for matching. Estimation then centres on the number of families with one or more victims; and survey information from survivors may be used to assess the demography or relatedness of victims, conditional upon the family having been victimized. Alternatively, the policy-relevant estimation goal may be the number of households in which at least one member is i) a victim of crime or ii) a perpetrator of crime(s), as interventions may target households rather than individuals.

6 Discussion

The estimation of hidden, or difficult to observe, populations will continue to be of interest across a wide spectrum of applications. The implications of such populations cut across many areas of public policy, from ecology through health and economics to criminal justice. Reliable and accurate estimates not only provide supporting evidence for the introduction or maintenance of policies but also in the assessment of the impact of policies. Furthermore, interrogation of well-fitting statistical models provides additional information on the underlying relationships between sources and/or additional covariates. This may provide further insight into the effectiveness of policies and/or suggest hypotheses about how individuals interact with the different lists, which may be different within covariate-strata.

Advances in data collection procedures, electronic recording of a wide variety of data and linkage methods provide increased potential for MSEs: by incorporating additional lists within the process or more individual-level information within the statistical analysis. However, this increases the ethical challenge of combining different data-sources, while maintaining data protection and data privacy.

Combining information from multiple sources is a powerful tool and there is a need to understand each different source, both in isolation and collectively with the other sources, to model accurately and incorporate any necessary complexities that may arise from the source or how it is combined with the other lists. These issues become increasingly important as the collection and storage of data become increasingly more automated and readily available for cross-classification. Increasingly fine detail may also become available. In short, data collection processes are accelerating at a faster pace than are the associated record linkage and statistical techniques for analysing such data.

The development of statistical tools to adapt to new challenges (as above) should be conducted in close collaboration between developers, data collectors/providers and policy makers to ensure that the importance of different factors is clearly understood and accounted for. In addition, this will assist policy makers in having an appreciation (ideally, an understanding) of the complexities and fine intricacies involved but also of the limitations associated with such studies. Understanding the limitations of MSE studies minimises the potential for their misuse through ignorance. Further, identifying such limitations may itself provide intuition and guidance for how data collection or presentation may be improved to answer more specific questions of interest (e.g. geographical level; different drug types; or species). This in turn may necessitate further statistical developments to answer the new questions of interest. Additionally, the results obtained from MSE studies need to be presented in an interpretable and flexible manner to allow for both focus on the effectiveness of “top-down” policies and the identification of “bottom-up” factors to understand better the underlying system and inter-relationships.

The use of multiple types of data can be a powerful tool for combining information. Such integrated approaches permit further insight into the hidden population and associated policies and impact; or into conflicting data (Prevost *et al*, 2015). For example, combining estimated population sizes of opioid users with data on the number of opioid related deaths permits the estimation of an opioid death rate per 100 users rather than per 1 million of population. In such calculations it is important to properly take into account uncertainty with regard to the estimated total numbers of opioid users and deaths. This again emphasises the importance of the close relationship between the statistical analyst and policy makers to ensure that the necessary output is available from the analysis to construct such estimates and that policy makers do not incorrectly calculate estimates through imperfect understanding of the statistical analysis.

The ethical case for MSE studies as a public good is strengthened by salient case-histories which demonstrate the impact of MSE “discoveries” on i) public policy (for example, about harm reduction, conservation, criminal justice), ii) revision of research agendas (to include validation studies or improve data-collection processes), or iii) performance monitoring (by national statistics or for resource allocation).

The Open Access versus deductive disclosure dilemma that MSE and record-linkage studies pose is addressed by safe-havens and the need for independent PAC-approvals in terms of mitigation against deductive disclosure. However, authors need indulgence from journal editors if their exposition of MSE methods is to be detailed enough to enable other research teams to apply the same criteria in other jurisdictions. Alternatively, MSE-teams may need to offer, as Open Access, their PAC application so that others can obtain the same data access as the original MSE-team was afforded. However, unless the originally-accessed linked-data are stored, application of the same matching criteria to updated source-files will not retrieve the original data-sets, see White *et al* (2017).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank David Borchers for useful discussions of some of the issues within the paper. SMB would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme on Probability and Statistics in Forensic Science where work on this paper was undertaken.

This work was supported by EPSRC grant no EP/K032208/1. RK was part-funded by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

Acronyms

HCV	Hepatitis C virus
MSE	Multiple systems estimation
OST	Opioid substitution therapy
PDU	Problem drug user

References

Baillargeon S, Rivest LP. Rcapture: Loglinear models for capture-recapture in R. J Stat Soft. 2007; 19:5.

Ball P, , Asher J, , Sulmont D, , 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 Washington, DC: Report to the Peruvian Commission for Truth and Justice (CVR); 2003

Besbeas P, , Borysiewicz RS, , Morgan BJT. Completing the ecological jigsawModeling Demographic Processes in Marked Populations Thomson DL, Cooch EG, , Conroy MJ, editorsNew York: Springer; 2009 51339

Bird SM, Leigh Brown AL. Criminalisation of HIV transmission: implications for public health in Scotland. BMJ. 2001; 323:1174–1177. [PubMed: 11711413]

Bird SM, Hutchinson SJ. Male drugs-related deaths in the fortnight after release from prison: Scotland, 1996-1999. Addiction. 2003; 98:185–190. [PubMed: 12534423]

Bird SM. Fatal accident inquiries into 97 deaths in prison custody in Scotland (1999-2003, or during first five years of operation of Scotlands only private prison): elapsed time to end of inquiry or written determination, issues and recommendations. Howard J Crim Just. 2008; 47:343–370.

Bird SM, Fairweather CB. Improvised explosive devices and military fatalities in Iraq and Afghanistan. J R Un Serv Inst. 2009; 154:30–38.

Bird SM, McAuley A, Munro A, Hutchinson SJ, Taylor A. Prison-based prescription of take-home naloxone for persons who inject drugs contributes to the effectiveness of Scotland's National Naloxone Policy, 2011-2015. Lancet. 2017; 389:1005–1006. [PubMed: 28290986]

Borchers DL, Efford MG. Spatially explicit maximum likelihood methods for capture-recapture studies. Biometrics. 2008; 64:377–85. [PubMed: 17970815]

Borchers DL, Distiller G, Foster R, Harmsen B, Milazzo L. Continuous-time spatially explicit capture-recapture, with an application to a jaguar camera- trap survey. Meth Ec Ev. 2014; 5:565–665.

Borchers DL, Fewster R. Spatial capture-recapture models. Stat Sci. 2016; 31:219–232.

Brooks SP, King R, Morgan BJT. A Bayesian approach to combining animal abundance and demographic data. An Bio Cons. 2004; 27:515–529.

Chapman DG. Some properties of the hypergeometric distribution with applications to zoological sample censuses. U of Cal Pub Stat. 1951; 1:131–160.

Cormack RM. Estimates of survival from the sighting of marked animals. Biometrika. 1964; 51:429–438.

Coull B, Agresti A. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. Biometrics. 1999; 55:294–301. [PubMed: 11318172]

Deborah Lader, editorCrime Survey for England and WalesStatistical Bulletin 7/16 second edition. Drug Misuse Findings from 2015/16Home Office; 2016 Jul. 2016

Darroch JN. The multiple recapture census. I. Estimation of a closed population. Biometrika. 1958; 45:343–59.

Durban JW, Elston DA. Mark-recapture with occasion and individual effects: Abundance estimation through Bayesian model selection in a fixed dimensional parameter space. JABES. 2005; 10:291–305.

Efford MG. Density estimation in live-trapping studies. Oikos. 2004; 106:598–610.

Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika. 1972; 59:591–603.

Fienberg S, Johnson M, Junker B. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. J R Stat Soc A. 1999; 163:383–405.

Fienberg SE, Manrique-Vallier D. Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. Adv Stat An. 2009; 93:49–60.

Fisher N, Turner SW, Pugh R, Taylor C. Estimated numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. BMJ. 1994; 308:27–30. [PubMed: 8298348]

Forster JJ, Gill RC, Overstall AM. Reversible jump methods for generalised linear models and generalised linear mixed models. Stats Comp. 2012; 22:107–120.

Gao L, Dimitropoulou P, Robertson JR, McTaggart S, Bennie M, Bird SM. Risk-factors for methadone-specific deaths in Scotland;s methadone-prescription clients between 2009 and 2013. Drug Alc Dep. 2016; 167:214–223.

Gimenez O, Choquet R. Incorporating individual heterogeneity in studies on marked animals using numerical integration: capture-recapture mixed models. Ecology. 2010; 91:951–57. [PubMed: 20462110]

Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974; 61:215–231.

Gore SM, Bird AG, Burns SM, Goldberg DJ, Ross AJ, Macgregor J. Drug injection and HIV prevalence in inmates of Glenochil prison. BMJ. 1995; 310:293–296. [PubMed: 7866170]

Goudie IBJ, Goudie M. Who captures the marks for the Petersen estimator? J Roy Stat Soc A. 2007; 170:825–839.

Grigg DB. Population Growth and Agrarian Change: An Historical Perspective Cambridge: Cambridge University Press; 1980 340

Hald A. A History of Probability and Statistics and Their Applications before 1750 New York: Wiley; 1990 586

Harron K, Goldstein H, , Dibben C, editorsMethodological Developments in Data Linkage Chichester: John Wiley & Sons; 2016 296

Hay G, , Gannon M, , Casey J, , McKeganey N. Estimating the national and local prevalence of problem drug misusers in Scotland Technical Report - University of Glasgow; 2009 2009 http://www.scotpho.org.uk/downloads/drugs/Prevalence_Report_%202006.pdf

Health Protection Scotland, University of the West of Scotland, Glasgow Caledonian University and the West of Scotland Specialist Virology CentreThe Needle Exchange Surveillance Initiative: Prevalence of blood-borne viruses and injecting risk behaviours among people who inject drugs attending injecting equipment provision services in Scotland, 2008-09 to 2015-16 Glasgow: Health Protection Scotland; 2017 Mar.

Hutchinson SJ, Goldberg DJ, Gore SM, Cameron S, McGregor J, McMenamin J, McGavigan J. Hepatitis B outbreak at Glenochil Prison during January to June 1993. Ep Inf. 1998; 121:185–191.

Hutchinson SJ. Modelling the hepatitis C virus disease burden among injecting drug users in Scotland University of Glasgow PhD Thesis; 2004

Hutchinson SJ, Bird SM, Goldberg DJ. Modelling the current and future disease burden of Hepatitis C among injecting drug users in Scotland. Hepatology. 2005; 42:711–723. [PubMed: 16116637]

Information Services Division, Scotland. Estimating the National and Local Prevalence of Problem Drug Use in Scotland 2012/13. 2016 https://isdscotland.scot.nhs.uk/Health-Topics/Drugs-and-Alcohol-Misuse/Publications/2014-10-28/2014-10-28-Drug-Prevalence-Report.pdf?33819216490

Jolly GM. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. Biometrika. 1965; 52:225–47. [PubMed: 14341276]

Jones HE, Hickman M, Welton NJ, De Angelis D, Harris RJ, Ades AE. Recapture or precapture? Fallibility of standard capture-recapture methods in the presence of referrals between sources. Am J Ep. 2014; 179:1383–93.

King R, Brooks SP. On the Bayesian analysis of population size. Biometrika. 2001a; 88:317–336.

King R, Brooks SP. Prior induction in log-linear modelling. Ann Stats. 2001b; 29:715–747.

King R, Bird SM, Brooks SP, Hutchinson SJ, Hay G. Prior information in behavioural capture-recapture methods: demographic influences on drug injectors' propensity to be listed in data sources and their drugs-related mortality. Am J of Ep. 2005; 162:1–10.

King R, Brooks SP. On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. Biometrics. 2008; 64:816–824. [PubMed: 18047534]

King R, Bird SM, Hay G, Hutchinson SJ. Updated estimation of the prevalence of injecting drug-users in Scotland via capture-recapture methods. Stat Meth Med Res. 2009a; 18:341–359.

King R, , Morgan BJT, , Gimenez O, , Brooks SP. Bayesian Analysis for Population Ecology Boca Raton: CRC Press; 2009b 456

King R, Bird SM, Overstall A, Hay G, Hutchinson SJ. Injecting drug users in Scotland, 2006: listing, number, demography, and opiate-related death-rates. Add Res Th. 2013; 21:235–246.

King R. Statistical ecology. Ann Rev Stat Appl. 2014; 1:401–426.

King R, Bird SM, Overstall A, Hay G, Hutchinson SJ. Estimating prevalence of injecting drug users and associated heroin-related death-rates in England using regional data and incorporating prior information. J R Stat Soc A. 2014; 177:1–28.

King R, McClintock B, Kidney D, Borchers DL. Abundance estimation using a semi-complete data likelihood approach. Ann Appl Stat. 2016; 10:264–285.

Knuiman MW, Speed TP. Incorporating prior information into the analysis of contingency tables. Biometrics. 1988; 44:1061–1071. [PubMed: 3233246]

Laska EM, Meisner M. A plant-capture method for estimating the size of a population from a single sample. Biometrics. 1993; 49:209–220. [PubMed: 8513102]

Lee A. Effect of list errors on the estimation of population size. Biometrics. 2002; 58:185–191. [PubMed: 11890314]

Lincoln FC. Calculating waterfowl abundance on the basis of banding returns. US Dept of Agri Circ. 1930; 118:1–4.

Madigan D, York JC. Bayesian methods for estimation of the size of a closed population. Biometrika. 1997; 84:19–31.

Manly BFJ, , McDonald TL, , Amstrup SC. Introduction to the HandbookHandbook of Capture-Recapture Analysis Amstrup SC, McDonald TL, , Manly BFJ, editorsNew Jersey: Princeton University Press; 2005 121

McCrea R, , Morgan B. Analysis of Capture-recapture Data Chapman and Hall/CRC Press; 2014 314

Merrall ELC, Kariminia A, Binswanger IA, Hobbs M, Farrell M, Marsden J, Hutchinson SJ, Bird SM. Meta-analysis of drug-related deaths soon after release from prison. Addiction. 2010; 105:1545–1554. [PubMed: 20579009]

Merrall ELC, Bird SM, Hutchinson SJ. Mortality of those who attended drug services in Scotland 1996-2006: record linkage study. Int J Drug Policy. 2012; 23:24–32. [PubMed: 21719267]

Millar T, , McAuley A. European Monitoring Centre for Drugs and Drug Addiction: Assessment of drug induced deaths data and contextual information in selected countries Lisbon: European Monitoring Centre for Drugs and Drug Addiction; 2017 in press

National Records of Scotland. [accessed 4 April 2017] Drug-related deaths in Scotland in 2015. 2016 https://www.nrscotland.gov.uk/files//statistics/drug-related-deaths/15/drugs-related-deaths-2015.pdf

Otis DL, Burnham KP, White GC, Anderson DR. Statistical inference from capture data on closed animal populations. Wild Mono. 1978; 62:1–135.

Overstall AM, King R. A Default Prior Distribution for Contingency tables with dependent factor levels. Stat Meth. 2014a; 16:90–99.

Overstall AM, King R. conting: an R package for Bayesian analysis of complete and incomplete contingency tables. J Stat Soft. 2014b; 58:1–27.

Overstall A, King R, Bird SM, Hay G, Hutchinson SJ. Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Stat Med. 2014; 33:1564–1579. [PubMed: 24293386]

Petersen CGJ. Report of the Danish Biological Station (1895). 1896; 6:5–84.

Pierce M, Bird SM, Hickman M, Millar T. National record-linkage study of mortality for a large cohort of opiate users ascertained by drug treatment or criminal justice sources, 2005-2009. Drug Alc Dep. 2015; 146:17–23.

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, Jones A, Millar T. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. Addiction. 2016; 111:298–308. [PubMed: 26452239]

Pierce M, , Millar T, , Robertson JR, , Bird SM. Ageing opioid users increased risk of methadone-specific death in the UK: irrespective of gender MRC Biostatistics Unit Technical Report; 2017

Pledger S. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics. 2000; 56:434–42. [PubMed: 10877301]

Prevost TC, Presanis AM, Taylor A, Goldberg DJ, Hutchinson SJ, de Angelis D. Estimating the number of people with hepatitis C virus who have ever injected drugs and have yet to be diagnosed: an evidence synthesis approach for Scotland. Addiction. 2015; 110:1287–1300. [PubMed: 25876667]

Reynolds TJ, King R, Harwood J, Frederikesen M, Harris MP, Wanless S. Integrated data analyses in the presence of emigration and tag-loss. JABES. 2009; 14:411–431.

Royle JA, , Chandler RB, , Sollmann R, , Gardner B. Spatial Capture-Recapture Academic Press; 2014 612

Sandland RL, Cormack RM. Statistical inference for Poisson and Multinomial models for capture-recapture experiments. Biometrika. 1984; 71:27–33.

Schnabel ZE. The estimation of total fish populations of a lake. Am Math Monthly. 1938; 45:348–352.

Seaman SR, Brettle RP, Gore SM. Mortality from overdose among injecting drug users recently released from prison: database linkage study. BMJ. 1998; 316:426–428. [PubMed: 9492665]

Seber GAF. A note on the multiple-recapture census. Biometrika. 1965; 52:249–59. [PubMed: 14341277]

Seybolt TB, Aronson JD, , Fischhoff B, editorsCounting Civilian Casualties: An Introduction to Recording and Estimating Non-military Deaths in Conflict Oxford: Oxford University Press; 2003 336

Silverman B. Modern slavery: An application of multiple systems estimation Home Office; 2014 Dec. 2014 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/386841/Modern_Slavery_an_application_of_MSE_revised.pdf

Spiegel PB, Salama P. War and mortality in Kosovo, 1998-99: an epidemiological testimony. Lancet. 2000; 355:2204–2209. [PubMed: 10881894]

Strang J, Hall W, Hickman M, Bird SM. Impact of supervision of methadone consumption on deaths related to methadone overdose (1993-2008): analyses using OD4 index in England and Scotland. BMJ. 2010; 341:c4851. [PubMed: 20847018]

Sutherland J, Schwarz CJ. Multi-list methods using incomplete lists in closed populations. Biometrics. 2005; 61:134–140. [PubMed: 15737086]

Sutherland J, Schwarz CJ, Rivest LP. Multilist population estimation with incomplete and partial stratification. Biometrics. 2007; 63:910–916. [PubMed: 17825020]

Taylor A, Goldberg D, Emslie J, Wrench J, Gruer L, Cameron S, Black J, Davis B, McGregor J, Follett E. Outbreak of HIV infection in a Scottish prison. BMJ. 1995; 310:289–292. [PubMed: 7866169]

Tilling K, Sterne JA. Capture-recapture models including covariate effects. Am J of Ep. 1999; 149:392–400.

UK Advisory Council on the Misuse of Drugs (chairman: Professor Sir Michael Rawlins)Reducing Drug-Related Deaths. A report by the Advisory Council on the Misuse of Drugs London: The Stationery Office; 2000

White SR, Bird SM, Grieve R. Review of methodological issues in cost-effectiveness analyses relating to injecting drug users, and case-study illustrations. J R Stat Soc A. 2014; 177:625–642.

White SJ, Bird SM, Merrall ELC, Hutchinson SJ. Drugs-related death soon after hospital-discharge among drug treatment clients in Scotland: record linkage, validation and investigation of risk-factors. Plos One. 2015; 10:e0141073. [PubMed: 26539701]

White SR, Muniz-Terrera G, Matthews FE. Sample size and classification error for Bayesian change-point models with unlabelled sub-groups and incomplete follow-up. Stat Meth Med Res. 2017 in press.

Worthington H, , McCrea RS, , King R. Estimation of population size when capture probability depends on individual states University of St Andrews Technical Report; 2017

Wright JA, Barker RJ, Schofield MR, Frantz AC, Byrom AE, Gleeson DM. Incorporating genotyping uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics. 2009; 65:833–40. [PubMed: 19173702]

Annotated references

Fienberg. Provides the foundation of log-linear models applied to contingency table data. 1972

Fisher, et al. Use of multiple systems estimation for additional hidden population (homeless). 1994

Jones, et al. Identification and impact of referrals within multiple systems estimation. 2014

King, Brooks. Describes a Bayesian model-averaging approach for hierarchical log-linear models. 2001a

King, et al. Validation of gender and age-group interaction for people who inject drugs. 2014

Overstall, King. Provides an R package for conducting Bayesian analysis of hierarchical log-linear models in the presence of model uncertainty. 2014b

Petersen. Lays the formal foundation of multiple systems estimation. 1896

Seybolt, et al. Provides guidance for multiple systems estimation and machine learning for rigorous reproducible matching. 2003

Silverman. Governmental report using multiple systems estimation for modern hidden populations (modern day slavery). 2014

Spiegel, Salama. Application of multiple systems estimation used within war crimes tribunal to corroborate evidence. 2000

Summary points

1. Within multiple systems estimation (MSE) approaches it is important to maintain a close relationship between data collectors, statistical analysts and policy makers in order to have a smooth transition from understanding the different sources of data, incorporating important factors into the analysis and interpreting output correctly (including at a range of different levels within a consistent manner).

2. In any statistical analysis there needs to be an understanding of the limitations of the approaches, including potentially large uncertainty, multi-modality and validation of MSE discoveries.

3. There is a potential tension between Open Access and risk of Deductive Disclosure in detailed MSE studies that should be understood prior to analyses and publication of results; where possible, data should be made available for reproducibility.

4. There are now numerous case-histories where MSE discoveries have altered policy and/or research agendas, nationally and internationally. MSE can act in the public good, bringing equity by counting, and thereby illuminating, the hard-to-reach and their plights.

Future issues

1. Multiple systems estimation (MSE) focuses only on a summary of the available information from the data in terms of the combination of sources an individual is observed by. This discards any temporal information, in terms of the exact times an individual is observed by each source and multiple recordings of an individual by a source. The use of such extended data will permit more intricate detail within the statistical modelling, providing further insight into an individual's trajectory through the sources and the potential for more insightful understanding of the system and population estimates.

2. How can presence/absence data for multiple systems estimation be combined with other additional data-sources or different forms of data to provide improved estimation and a greater understanding of the underlying system? This includes addressing additional issues, for example, different sources using different identifying information, non-unique identifiers, unknown sub-strata for some individuals and ensuring data privacy is maintained.

3. With increasingly advanced statistical techniques and associated computational power, it will become even more important to create accessible computer packages and associated training to permit use of advanced techniques within government, charities etc. and an understanding of the interpretation of the output, including associated limitations. This includes understanding the best ways to present the results of the statistical analyses to policy makers in a comprehensive yet clear and interpretable manner which includes the quantification of the uncertainties associated with any estimates.

Table 1

A 2^3 incomplete contingency table, where S_i denotes survey $i = 1, \dots, 3$ and n_{000} is unobserved and hence is denoted by a “?”.

		$S_1 = 0$	$S_1 = 1$
$S_3 = 0$	$S_2 = 0$	?	n_{100}
	$S_2 = 1$	n_{010}	n_{110}
$S_3 = 1$	$S_2 = 0$	n_{001}	n_{101}
	$S_2 = 1$	n_{011}	n_{111}