ANNUAL REVIEWS

Annual Review of Statistics and Its Application

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

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

[1]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SR, United Kingdom; email: sheila.bird@mrc-bsu.cam.ac.uk

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

[3]School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom; email: ruth.king@ed.ac.uk

Annu. Rev. Stat. Appl. 2018. 5:95–118

First published as a Review in Advance on December 13, 2017

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031017-100641

Keywords

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

Abstract

Applications of estimating population sizes range from estimating human or ecological population size within regions or countries to estimating the hidden number of civilian casualties in war. Total enumeration via a census is typically infeasible. However, a series of partial enumerations of a population is often possible, leading to capture-recapture methods, which have been extensively used in 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 multiple lists, often referred to as multiple systems estimation (MSE). The MSE approach is preferred when estimating capture-shy or hard-to-reach populations, including those who are caught up in the criminal justice system, trafficked, or civilian casualties of war. Motivated by the public policy applications of MSE, each briefly introduced, we discuss practical problems with methodological implications. They include period definition; case definition; scenarios 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 lists; demographic or other information about the case that influences capture propensities; permissions to access 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 finally on whether the uncertainty that qualified their estimates was sufficiently narrow to orient public policy.

1. BRIEF HISTORY OF MULTIPLE SYSTEMS (OR CAPTURE-RECAPTURE) ESTIMATION

We briefly outline the history of multiple systems (or capture-recapture) estimation (MSE) of population sizes. The approach has been applied in numerous areas, from wildlife populations (King & 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). A video of another demonstration is available at **https://www.youtube.com/watch?v=aiSKgIc_8vk**.

The idea of combining information from two different partial enumerations has a long history dating 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 (*a*) birth records of babies born across the whole of France, and (*b*) 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:

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

thus 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) (**Table 1**).

In **Table 1**, the level of 0 corresponds to not being observed by the given survey, and 1 corresponds to 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 (*a*) the proportion of individuals recorded by the municipality censuses that are also recorded on the birth certificates, and (*b*) the proportion of individuals in the total population that are also recorded on the birth certificates. Assuming that the two partial enumeration processes

MSE: multiple systems estimation

Table 1 2×2 **incomplete contingency table**

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

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 the national birth certificates and the census of the given municipalities, respectively, and let n_{11} denote the number of individuals observed by both surveys. Finally let N denote the total number of individuals. Applying the above rationale we obtain an estimate of N, denoted $\hat{N}$, as follows:

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

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) nonparametric 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 in fisheries (Petersen 1896, Lincoln 1930). For a history of this estimator, see Goudie & 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 following less-biased estimator for the total population size:

$$\hat{N} = \frac{(n_{1.}+1)(n_{.1}+1)}{(n_{11}+1)}.$$

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 in the study, detailing whether or not the individual was observed by each survey. The data are usefully summarized in the form of an incomplete contingency table corresponding to the number of individuals observed by each distinct combination of surveys.

Owing 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

Table 2 A 2^3 incomplete contingency table, where S_i denotes survey $i = 1, \ldots, 3$, and n_{000} is unobserved and hence is denoted by a question mark

		$S_1 = 0$	$S_1 = 1$
$S_3 = 0$	$S_2 = 0$	?	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}

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 (such as a tag or ring) is applied that can be uniquely identified at subsequent capture occasions; more recently, photographic identification may be used rather than an applied mark. In 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, or 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 or 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, which are often conducted over a longer period of time, the assumption of closure (no births/deaths/migration) was relaxed, leading to the Cormack-Jolly-Seber model (Cormack 1964, Jolly 1965, Seber 1965); King (2014) provides 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 and permitting the inclusion of a trap-happy or trap-shy response following initial capture, i.e., behavioral effects) and individual heterogeneity. These dependencies are usefully summarized and described by Otis et al. (1978). Detailed discussions of these models and associated extensions are provided by, for example, King (2014) and McCrea & 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 subpopulations (strata), such that each subpopulation (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 MSE 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 ith digit of each triple corresponds to being observed (=1) or not observed (=0) by survey i. We let n_{ijk} denote the number of individuals observed by the combination of sources denoted by $ijk \in R \cup \{000\}$. We can again represent the data in an incomplete contingency table (**Table 2**).

We assume that the cell counts are independent, conditional on the model parameters, such that, given expected counts μ_{ijk},

$$n_{ijk}|\mu_{ijk} \sim \text{Poisson}(\mu_{ijk}),$$

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 $(\theta_i^1, \theta_j^2, \theta_k^3)$ reflect the propensity of being observed by each source, and the two-way interactions $(\theta_{ij}^{12}, \theta_{ik}^{13}, \theta_{jk}^{23})$ reflect the dependence between each survey pair. This is the saturated model because, 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 maximum likelihood estimates). Submodels 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 or negative) can also be of interest for public policy to understand how individuals in the population interact with the different surveys. Sandland & 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 or adapted to allow for additional factors. For example, Tilling & Sterne (1999) discussed 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 analyzed independently, although this can lead to relatively few overlaps between sources (meaning 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 addressed this issue by analyzing all strata simultaneously, permitting the borrowing of information across the strata by specifying an extended log-linear model and the treatment of 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, which permit both the inclusion of prior information on the total population size and/or log-linear terms, and Bayesian model-averaging can incorporate both parameter and model uncertainty (Knuiman & Speed 1988; Madigan & York 1997; Fienberg et al. 1999; King & Brooks 2001a,b; King et al. 2005; Overstall & King 2014a,b; King 2014).

2. APPLICATIONS OF MULTIPLE SYSTEMS ESTIMATION

We introduce six applications of MSE that have policy implications ranging from the environment, to public health and criminal justice, to human rights. 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, for example by providing denominators en route to additional calculations, say, of death rates, 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 for 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 the use of a range of different data collection techniques, 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 have 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 and management purposes in order to predict effects related to, for example, changes in food availability and/or habitat.

OST: opioid substitution therapy

HCV: hepatitis C virus

PDU: problem drug user

2.2. Persons Who Inject Heroin

Heroin injectors incur criminal justice, welfare, and health care 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 (Seaman et al. 1998; Bird & Hutchinson 2003; Merrall et al. 2010, 2012; King et al. 2013, 2014; Pierce et al. 2015; White et al. 2015). Injectors are difficult to count, 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), in prison, or hospitalized on account of a nonfatal overdose, a mental health problem, an external injury, a blood-borne virus or other infectious disease, 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-behavior has decreased, and Scotland's HCV Action Plans have successfully promoted confidential HCV testing of persons born in 1956–1975 who have ever injected, as a high proportion are expected to be HCV-antibody positive (Hutchinson et al. 2005). However, this promotion has led to further issues as, increasingly, those who declare injecting as their HCV risk-behavior are former, not current, injectors (see Section 3.4 for further discussion).

2.3. Problem Drug Users

Case definitions of problem drug users (PDUs) can 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 (Lader 2016), which includes computerized questions on past-year use of illegal drugs, have been considered as alternatives to MSE for estimation of PDUs. However, household-based surveys seriously underestimate past-year use of hard drugs such as heroin because notable proportions 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 offenses. The policy's intention was the referral of those testing positive

to drug treatment agencies. However, Jones et al. (2014) showed that the more successful the referrals are at engaging individuals who test positive into drug treatment, the more problematic for MSE, as the 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 interdependent (in one direction).

2.4. Homeless Individuals

Censuses of homeless individuals in a defined district and period (say, midnight to 3 AM on the census day) typically lead to an undercount of the true population. For example, volunteers who enumerate the number of individuals within a given region and time period at hostels or on the streets typically miss homeless people who are temporarily accommodated (e.g., in custody or hospitalized). Furthermore, the volunteers cannot ascertain outdoor sleepers' age-groups or other demographic data without waking 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, which rely on a single capture occasion in a given area and time (Laska & Meisner 1993), have also been applied in the United States. In such a study, volunteers are planted in the community and report whether they have been observed during the single survey, leading to the application of a dual system 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, incentivized, 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, e.g., **https://www.theguardian.com/uk/2011/feb/06/sex-traffick-romania-britain**). It is a matter of controversy whether, in the wider public interest, the victims of human trafficking should be excused from prosecution for serious crimes, ranging from drug trafficking up to manslaughter, that they have been forced, or driven, to commit (see **http://www.carmelitechambers.co.uk/news-and-events/news/human-trafficking-victims-should-not-be-charged-with-murder-felicity-gerry**). If victims are not given immunity from prosecution, this may further inhibit them 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 that enumerated 2,744 potential victims of trafficking into the United Kingdom and estimated a further 7,000 to 10,000 as the unenumerated hidden figure.

2.6. Crimes Against Humanity

The work of the Human Rights Data Analysis Group (HRDAG), which has now existed for more than 25 years, has featured in high-profile legal cases such as the conviction in Guatemala of General E.R. Montt for genocide and crimes against humanity and has underpinned truth and reconciliation commissions. In Peru, for example, HRDAG estimated that approximately 70,000 deaths had occurred in the period of the Commission's purview, of which only 25,000 had been

documented directly (18,000 by the Commission, another 7,000 by data sources other than the Commission).

Seybolt et al. (2003) described 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 versus 2), the next step is to synchronize coding across relevant datasets. Alignment may lead to some degradation if the datasets have recorded information differently (e.g., child/adult or <15 years/15–44 years/45–64 years/65+ years). If there has been no preagreement 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, and lethal force might all map to "homicide." It is far better if a common vocabulary can be agreed on in advance and used by each data source.

Stratification of datasets, prior to identifying overlaps, is often prudent. For example, whether a homicide is listed by specific data sources may depend on characteristics of the victim, such as sex and age-group, or on the geographical area in which it happened, for example, during conflicts such as those in Syria or Afghanistan (Bird & Fairweather 2009). Overstratification can, of course, result in too few overlaps for MSE to be sensibly applied.

3. RECENT METHODOLOGICAL DEVELOPMENTS

3.1. Statistical Ecology

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

Advances in technology have led to new issues and statistical solutions. For example, the identification of individuals using DNA matching from hair or scat samples or using photographic recognition have different potentials for mismatching that need to be accounted for in the statistical analysis to avoid bias being introduced (Wright et al. 2009). Alternatively, acoustic recordings may be used to identify individuals, rather than physical resightings and identifiable characteristics. For difficult-to-observe populations, an array of motion-sensitive camera or acoustic traps may be erected that captures animals passing by. This spatial information has led to the development of spatially explicit capture-recapture modeling (Efford 2004, Borchers & Efford 2008, Royle et al. 2014) and continuous time observations (Borchers et al. 2014). For a review, see, for example, Borchers & Fewster (2016).

3.2. Period Definition

The definition of the period for which the number of prevalent current injectors (say) needs to be estimated, typically in a calendar year, balances the time required for there to be a reasonable chance of "listing" the persons of interest by the various MSE data sources with timeliness for the prevalence-estimation to be relevant to policy. Thus, for example, Scotland estimated its calendar-year numbers 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, as well as the extent to which older injectors were (or were not) aging out of injecting, or dying from overdose.

Injecting is a relapsing-remitting behavior. Thus, of further MSE interest is estimation of how many individuals who were listed as injectors in 2008 are persistent in the sense of being also listed in 2011. Matching of MSE individuals across (and not just within) calendar-year 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 that compiles and assesses the overlaps between lists in 2008 may, of course, be different from the team funded to do so in 2011. Analysts typically receive only demographically prespecified overlap counts—for example, for each combination of region, sex, and age-group. Persistence can only be addressed by revisiting the fieldwork for both 2008 and 2011, and requires that both fieldwork teams have access to individual years of age to aid matching across periods, and also that clients did not migrate much from one region to another. Migration to live away from fellow injectors is, however, also a means to reduce clients' relapse into injecting.

3.3. Case Definition

Prior agreement on case definition is important so that each data source can determine, for its own list, who qualifies as a case and who does not. Scotland and England adopted different contemporary definitions for PDUs (Section 2.3). Moreover, Scotland's case definition for PDUs has evolved over time as the inclusion of persons in receipt of OST assumed greater importance than it had in the twentieth century when OST recipients were proportionately fewer (Strang et al. 2010). By the twenty-first century, optimism that OST meant cessation from injecting had given way to greater realism that OST reduces clients' frequency of injecting but does not mean that they necessarily cease injecting.

Case definition may also be less well adhered to by some data sources than others. For example, the risk-behavior 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, testing has been targeted more to 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-behavior, such as injecting, that 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. Thus, a decision has to be made as to whether case definition for injectors among Scotland's registered HCV diagnoses should include those whose risk-behavior is undeclared.

3.4. Observed Count as Upper Bound for the Count of Interest

If a data source cannot apply the case definition but can apply a broader definition, then the research team can elect to consider the observed count from that data source as an upper bound for the

count of interest if the subject was not recorded by any of the research team's other data sources. 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-behavior 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 modeled 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 a significant overestimate of the total number of current injectors (by a factor of two for 2009) and a potential misinterpretation that there was a constant population of injectors over time. The new methodology produced a progressive reduction in Scotland's estimated number of current injectors in concert with Scotland's HCV 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

One of the best-documented examples of robust specification of case matching across lists is when MSE is used to quantify crimes against humanity. As the intention is to bring the perpetrators of those crimes to justice, rigorous specification is essential because the court's judgment on capital crimes may rest upon the rigor 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. Rigor is essentially codified by how the matching program is written (see, e.g., 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, say A to D, so that according to some optimization criterion, the analysis can update—and indeed optimize—the match-weights in context as the estimation task and list propensities unfold (Harron et al. 2016). Although 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 & Schwarz (2005) extended 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 Cases Influences Capture Propensities

When list holders and analyst teams differ, the analysts have to specify in advance the cross-classified covariate strata. For example, the strata could be defined by sex (2 levels) × age-group

(3 levels) × geography (3 levels), in which case there are 18 strata, and for each of these, the 15 overlap counts across lists *A* to *D* (as in the example above) 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 recipients), *C* (child taken into 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 foster care, were receiving 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 receive can be disclosed. In some cases, the complete data may be made available to researchers under confidentiality agreements, but 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 an asterisk and, consequently, the 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 they make their judgments on the basis of the public interest cases 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 for transfer to a safe haven for matching across the prepared lists, to be programmed according to the PAC-approved rules that were defined by the analyst team.

3.8. List Creation by Research Teams or Interested Parties

Hay et al. (2009) made considerable efforts to assist local drug treatment teams in creating client lists that were suitable for use in separate MSEs of Scotland's number of injectors and 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 owing to convergence difficulties but also in recognition that OST clients, now counted as PDUs, might simultaneously qualify as currently injecting (ISD Scotl. 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 the 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 offenses 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 opiates/cocaine and drug treatment clients (see Jones et al. 2014).

3.10. Different Model Frameworks Giving Rise to Widely Different Estimations

King et al. (2013, 2014) found that limitation on the allowable interactions (first order, second order, third order) was more important in MSE than whether the analysis was conducted from a Bayesian or frequentist standpoint. The Bayesian framework has the advantage of allowing analysts (*a*) 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 (*b*) 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 that incorporate both model and parameter uncertainty (King & 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, because of inclusion or 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 (Overstall & King 2014a, King et al. 2005). 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 (DRD) 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 appropriate allowance for interactions is impossible owing 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 people 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 United Kingdom, MSE deduction about the age-related loss of female advantage in terms of DRDs 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 DRDs for some 33,000 methadone prescription clients in Scotland (Gao et al. 2016).

Pierce et al. (2016) investigated the impact on DRD rates for clients referred into drug treatment by the criminal justice system and found that the benefit of drug treatment was significantly reduced, but not negligible, for those clients.

3.12. Implementation of Multiple Systems Estimation 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. McCrea and Morgan (2014; section 2.13), for example, present 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 & Rivest 2007) fits a variety of models, calculating the associated maximum likelihood estimates 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 & King 2014b) provides posterior estimates for population sizes and interaction terms, allows for multiple covariates with a general number of levels and 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 PURPOSES?

Is the MSE uncertainty sufficiently narrow to orient public policy? If not, what options can be taken to reduce uncertainty or seek external validation? As importantly, is 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

In 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, number of chicks fledged, etc.), population counts, and dead recoveries. Many of these data sources provide information on the same model parameters, such as population size, survival probabilities, and fecundity rates. Analyzing 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 (potentially significantly) the precision of the parameter estimates (Brooks et al. 2004, Besbeas et al. 2009). In addition, comparing the integrated estimates with those obtained from each individual analysis may provide evidence of inconsistencies that can be investigated further, providing novel insight into the ecological system and allowing the model to be 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 because of, for example, 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 1999–2000 and ended in 2009–2010 when estimates were produced for PDUs but not for current injectors. For each such estimation, the MSE uncertainty was wide, but not so wide that one could not discern the following: 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 because 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 DRDs occurred in current injectors resulted in estimated DRD rates per 100 current injectors that 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 DRD rate) for Scotland's clients is substantially higher than for their counterparts in England (Pierce et al. 2017).

Perhaps unsurprisingly, Scotland's more recent MSEs have focused on PDUs including 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. Second, the MSE that was attempted in 2009–2010 and that used just three centrally available electronic data sources either failed to converge or gave answers that the review group was not confident about. MSE of current injectors has not been reported for 2012–2013 (ISD Scotl. 2016).

As described in Section 3.11, the sex by age-group interaction for 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 multiparameter evidence synthesis to estimate Scotland's prevalence of HCV carriage in former injectors, but the hoped-for refinement of MSEs by incorporating the age-specificity of HCV diagnoses was not forthcoming because the HCV prevalence estimates were themselves sensitive to inclusion of the serial MSEs.

4.3. Problem Drug Users

The UK Advisory Council on the Misuse of Drugs (Advis. Counc. Misuse Drugs 2000) called for DRD rates to be reported per 100 PDUs rather than per 100,000 of population because PDU prevalence was not homogeneous across the United Kingdom. Scotland's MSEs (with uncertainty) for its PDUs have been in the range 55,000 to 60,000 for more than 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 Scotland's National Naloxone Programme so that, by the end of 2015–2016, naloxone kits had been supplied to 30% of each region's PDUs (Bird et al. 2017). The PDU estimates (with uncertainty) also feature in Scotland's official statistics on DRD rates per 100 PDUs, which are reported by (*a*) sex and age-group and (*b*) NHS region (Natl. Rec. Scotl. 2016). Finally, Millar & McAuley (2017), on behalf of the European Monitoring Centre for Drugs and Drug Addiction, were 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. The Homeless

In the United Kingdom, at least, central estimates (with uncertainty) for the number of homeless people in major cities have not hit the headlines. Nor has there been much discussion about other approaches. For example, 30% of nearly 2,000 current injectors interviewed in 2015–2016 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 public assistance claimants might pose the same question, and a patchwork of estimates, taking overlaps between respondents into account, might be stitched together. Charities or churches that 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 the United Kingdom'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 United Kingdom in the twenty-first 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 five-year period and by eliciting expert opinion on, for example, the plausible external-cause death rate that 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: 9,000 to 12,000) were killed during March to June 1999, a claim disputed by the defense at the International Tribunal for the Former Yugoslavia but corroborated by a survey-based estimate (see Spiegel & Salama 2000) of 12,000 fatalities (95% confidence interval: 5,000 to 18,000), most during March to June 1999. By 2011, in further corroboration of the MSE by Ball et al. (2003), the Kosovo Memory Book had documented 14,000 deaths and disappearances, some of them military, between January 1998 and December 1999, and 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 recognize that well-understood practicality must ultimately have a technical translation for there to be realistic modeling of complexity.

5.1. Limited Number of Estimation Goals

For human populations, MSE and record-linkage studies both typically require a study protocol that 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, and 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 three years, say, for a single species or client-type 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 that account for the three-year persistence, for example, of persons who inject drugs could be devised but typically would require that the research team have access to source lists from three 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 three-year persistence may require comparison between the archived source lists from three years ago and its current format as well as the present-day source list. Record-linkage studies that focus on event sequences may of course come to borrow ideas from MSE when accounting for 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. If the pseudonymized linked data (that is, across data sources *A*, *B*, *C*, *D*)

are accessible by any of the contributing data sources, an individual's event sequence on any of the data sources (say A) may be sufficiently detailed for that data source (A) to de-identify the person and thereby 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 United Kingdom has established safe havens in which the linked database can be analyzed but from which copies of the linked database can neither be removed nor made open access.

Deductive disclosure about individuals is also a risk in MSE studies when source lists are centralized 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 offenses committed by anonymized subjects who may, as a consequence, be liable to prosecution if a court order has been obtained that requires their identification. Similar concerns may apply in MSE if the research team holds sensitive lists of identifiable individuals.

A record-linkage study (Hutchinson et al. 1998) 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 diagnosed with HIV. And if so, the next step was to ascertain whether, on the basis of their HIV diagnostic sample, they had been infected with the same strain of HIV as had 13 of the 14 HIV seroconversions that were reported by an Infection Control Exercise at Glenochil Prison (Taylor et al. 1995). Approximately 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 & 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 same strain of HIV as Mr. Kelly.

Record-linkage research teams face sanctions in the United Kingdom 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 deaths in Scottish prison custody in the first five years of the twenty-first 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, and then was reincarcerated in Shotts Prison, where he was diagnosed with AIDS lymphoma and died in 2001. Bird (2008) had inadvertently discovered the fourteenth 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. The reader is directed to **https://hrdag.org/2013/03/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, say, within families, so that MSE may occasionally need to consider family versus individual members within a household as the unit for matching. Estimation then centers 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 (*a*) a victim of crime or (*b*) 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 provide supporting evidence not only 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 into the process or more individual-level information into 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 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 analyzing 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) not only of the complexities and fine intricacies involved, but also of the limitations associated with such studies. Understanding the limitations of MSE studies minimizes the potential for their misuse through ignorance. Furthermore, 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., by geographical level, for 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 interrelationships.

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 take into account uncertainty properly with regard to the estimated total numbers of opioid users and deaths. This again emphasizes the importance of the close relationship between statistical analysts 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 that demonstrate the impact of MSE discoveries on (*a*) public policy (for example, about harm reduction, conservation, or criminal justice), (*b*) revision of research agendas (to include validation studies or improve data collection processes), or (*c*) 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 their PAC application via open access 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 datasets (see White et al. 2017).

SUMMARY POINTS

1. In 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 in 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 in which MSE discoveries have altered policy and/or research agendas, both in the United Kingdom 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 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, such as 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 in the statistical modeling, 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 MSE 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 issues such as different sources using different identifying information, nonunique identifiers, unknown substrata for some individuals, and ensuring that 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 in groups such as the government and charities, 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 that includes the quantification of the uncertainties associated with any estimates.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to thank David Borchers for useful discussions of some of the issues in the article. S.M.B. would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the program 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. R.K. was part-funded by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

LITERATURE CITED

Advis. Counc. Misuse Drugs. 2000. *Reducing Drug-Related Deaths*. London: Stationery Off.

Baillargeon S, Rivest LP. 2007. `Rcapture`: Loglinear models for capture-recapture in R. *J. Stat. Softw.* 19(5):1–31

Ball P, Asher J, Sulmont D, Manrique D. 2003. *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: Am. Assoc. Adv. Sci.

Besbeas P, Borysiewicz RS, Morgan BJT. 2009. Completing the ecological jigsaw. In *Modeling Demographic Processes in Marked Populations*, ed. DL Thomson, EG Cooch, MJ Conroy, pp. 513–39. New York: Springer

Bird SM. 2008. Fatal accident inquiries into 97 deaths over five years in Scottish prison custody: long elapsed times and recommendations. *Howard J. Crim. Just.* 47:343–70

Bird SM, Fairweather CB. 2009. IEDs and military fatalities in Iraq and Afghanistan. *J. R. United Serv. Inst.* 154:30–38

Bird SM, Hutchinson SJ. 2003. Male drugs-related deaths in the fortnight after release from prison: Scotland, 1996–1999. *Addiction* 98:185–90

Bird SM, Leigh Brown AL. 2001. Criminalisation of HIV transmission: implications for public health in Scotland. *BMJ* 323:1174–77

Bird SM, McAuley A, Munro A, Hutchinson SJ, Taylor A. 2017. Prison-based prescriptions aid Scotland's National Naloxone Programme. *Lancet* 389:1005–6

Borchers DL, Distiller G, Foster R, Harmsen B, Milazzo L. 2014. Continuous-time spatially explicit capture-recapture models, with an application to a jaguar camera-trap survey. *Methods Ecol. Evol.* 5:656–65

Borchers DL, Efford MG. 2008. Spatially explicit maximum likelihood methods for capture-recapture studies. *Biometrics* 64:377–85

Borchers DL, Fewster R. 2016. Spatial capture-recapture models. *Stat. Sci.* 31:219–32

Brooks SP, King R, Morgan BJT. 2004. A Bayesian approach to combining animal abundance and demographic data. *Animal Biodivers. Conserv.* 27:515–29

Chapman DG. 1951. *Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses*. Berkeley: Univ. Calif. Press

Cormack RM. 1964. Estimates of survival from the sighting of marked animals. *Biometrika* 51:429–38

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

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

Durban JW, Elston DA. 2005. Mark: recapture with occasion and individual effects: abundance estimation through Bayesian model selection in a fixed dimensional parameter space. *J. Agric. Biol. Environ. Stat.* 10:291–305

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

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

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

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

Forster JJ, Gill RC, Overstall AM. 2012. Reversible jump methods for generalised linear models and generalised linear mixed models. *Stat. Comput.* 22:107–20

Gao L, Dimitropoulou P, Robertson JR, McTaggart S, Bennie M, Bird SM. 2016. Risk-factors for methadone-specific deaths in Scotland's methadone-prescription clients between 2009 and 2013. *Drug Alcohol. Depend.* 167:214–23

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

Goodman LA. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215–31

Gore SM, Bird AG, Burns SM, Goldberg DJ, Ross AJ, Macgregor J. 1995. Drug injection and HIV prevalence in inmates of Glenochil prison. *BMJ* 310:293–96

Goudie IBJ, Goudie M. 2007. Who captures the marks for the Petersen estimator? *J. R. Stat. Soc. A* 170:825–39

Grigg DB. 1980. *Population Growth and Agrarian Change: An Historical Perspective*. Cambridge, UK: Cambridge Univ. Press

Hald A. 1990. *A History of Probability and Statistics and Their Applications Before 1750*. New York: Wiley

Harron K, Goldstein H, Dibben C, eds. 2016. *Methodological Developments in Data Linkage*. Chichester, UK: Wiley

Hay G, Gannon M, Casey J, McKeganey N. 2009. *Estimating the national and local prevalence of problem drug misuse in Scotland*. Executive Rep., Univ. Glasgow. **http://www.scotpho.org.uk/downloads/drugs/Prevalence_Report_%202006.pdf**

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

Uses multiple systems estimation for additional hidden population (the homeless).

Health Protection Scotland. 2017. *The 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*. Rep., Health Protection Scotland, Glasgow, Scotl.

Hutchinson SJ. 2004. *Modelling the hepatitis C virus disease burden among injecting drug users in Scotland*. PhD Thesis, Univ. Glasgow

Hutchinson SJ, Bird SM, Goldberg DJ. 2005. Modeling the current and future disease burden of hepatitis C among injecting drug users in Scotland. *Hepatology* 42:711–23

Hutchinson SJ, Goldberg DJ, Gore SM, Cameron S, McGregor J, et al. 1998. Hepatitis B outbreak at Glenochil Prison during January to June 1993. *Epidemiol. Infect.* 121:185–91

ISD Scotl. (Inf. Serv. Div. Scotl.). 2016. *Estimating the national and local prevalence of problem drug use in Scotland 2012/13*. Publ. Rep., ISD Scotl. **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. 1965. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* 52:225–47

Identification and impact of referrals within multiple systems estimation.

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

King R. 2014. Statistical ecology. *Annu. Rev. Stat. Appl.* 1:401–26

King R, Bird SM, Brooks SP, Hutchinson SJ, Hay G. 2005. Prior information in behavioral capture-recapture methods: demographic influences on drug injectors' propensity to be listed in data sources and their drug-related mortality. *Am. J. Epidemiol.* 162:694–703

King R, Bird SM, Hay G, Hutchinson SJ. 2009a. Estimating current injectors in Scotland and their drug-related death rate by sex, region and age-group via Bayesian capture-recapture methods. *Stat. Methods Med. Res.* 18:341–59

Validation of gender and age-group interaction for people who inject drugs.

King R, Bird SM, Overstall A, Hay G, Hutchinson SJ. 2013. Injecting drug users in Scotland, 2006: listing, number, demography, and opiate-related death-rates. *Addict. Res. Theory* 21:235–46

King R, Bird SM, Overstall A, Hay G, Hutchinson SJ. 2014. Estimating prevalence of injecting drug users and associated heroin-related death-rates in England by using regional data and incorporating prior information. *J. R. Stat. Soc. A* 177:209–36

Describes a Bayesian table analysis model-averaging approach for hierarchical log-linear models.

King R, Brooks SP. 2001a. On the Bayesian analysis of population size. *Biometrika* 88:317–36

King R, Brooks SP. 2001b. Prior induction in log-linear models for general contingency. *Ann. Stats.* 29:715–47

King R, Brooks SP. 2008. On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics* 64:816–24

King R, McClintock B, Kidney D, Borchers DL. 2016. Capture-recapture abundance estimation using a semi-complete data likelihood approach. *Ann. Appl. Stat.* 10:264–85

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

Knuiman MW, Speed TP. 1988. Incorporating prior information into the analysis of contingency tables. *Biometrics* 44:1061–71

Lader D, ed. 2016. *Drug misuse: findings from the 2015 to 2016 CSEW*. Stat. Bull. 07/16, Home Off., U.K. 2nd ed.

Laska EM, Meisner M. 1993. A plant-capture method for estimating the size of a population from a single sample. *Biometrics* 49:209–20

Lee A. 2002. Effect of list errors on the estimation of population size. *Biometrics* 58:185–91

Lincoln FC. 1930. *Calculating waterfowl abundance on the basis of banding returns*. Circ. No. 118, USDA, Washington, DC

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

Manly BFJ, McDonald TL, Amstrup SC. 2005. Introduction to the Handbook. In *Handbook of Capture-Recapture Analysis*, ed. SC Amstrup, TL McDonald, BFJ Manly, pp. 1–21. New Jersey: Princeton Univ. Press

McCrea R, Morgan B. 2014. *Analysis of Capture-Recapture Data*. Boca Raton, FL: Chapman and Hall/CRC

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

Merrall ELC, Kariminia A, Binswanger IA, Hobbs M, Farrell M, et al. 2010. Meta-analysis of drug-related deaths soon after release from prison. *Addiction* 105:1545–54

Millar T, McAuley A. 2017. *EMCDDA assessment of drug-induced death data and contextual information in selected countries*. Tech. Rep., EMCDDA, Lisbon

Natl. Rec. Scotl. 2016. *Drug-related deaths in Scotland in 2015*. **https://www.nrscotland.gov.uk/files//statistics/drug-related-deaths/15/drugs-related-deaths-2015.pdf**

Otis DL, Burnham KP, White GC, Anderson DR. 1978. Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* 62:3–135

Overstall AM, King R. 2014a. A default prior distribution for contingency tables with dependent factor levels. *Stat. Methodol.* 16:90–99

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

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

Overstall A, King R, Bird SM, Hay G, Hutchinson SJ. 2014. Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Stat. Med.* 33:1564–79

Petersen CGJ. 1896. The yearly immigration of young plaice into the Limfjord from the German Sea. *Rep. Danish Biol. Stat. (1895)* 6:5–84

Lays the formal foundation of multiple systems estimation.

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, et al. 2016. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. *Addiction* 111:298–308

Pierce M, Bird SM, Hickman M, Millar T. 2015. National record-linkage study of mortality for a large cohort of opioid users ascertained by drug treatment or criminal justice sources in England, 2005–2009. *Drug Alcohol Depend.* 146:17–23

Pierce M, Millar T, Robertson JR, Bird SM. 2017. *Ageing opioid users' increased risk of methadone-specific death in the UK: irrespective of gender*. Tech. Rep., MRC Biostat. Unit. **https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/2014/02/SMB2017_1.pdf**

Pledger S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* 56:434–42

Prevost TC, Presanis AM, Taylor A, Goldberg DJ, Hutchinson SJ, de Angelis D. 2015. 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* 110:1287–300

Provides guidance for multiple systems estimation and machine learning for rigorous reproducible matching.

Reynolds TJ, King R, Harwood J, Frederikesen M, Harris MP, Wanless S. 2009. Integrated data analysis in the presence of emigration and mark loss. *J. Agric. Biol. Environ. Stat.* 14:411–31

Royle JA, Chandler RB, Sollmann R, Gardner B. 2014. *Spatial Capture-Recapture*. New York: Academic

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

Schnabel ZE. 1938. The estimation of total fish populations of a lake. *Am. Math. Mon.* 45:348–52

Seaman SR, Brettle RP, Gore SM. 1998. Mortality from overdose among injecting drug users recently released from prison: database linkage study. *BMJ* 316:426–28

Seber GAF. 1965. A note on the multiple-recapture census. *Biometrika* 52:249–59

Seybolt TB, Aronson JD, Fischhoff B, eds. 2003. *Counting Civilian Casualties: An Introduction to Recording and Estimating Non-Military Deaths in Conflict*. Oxford, UK: Oxford Univ. Press

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

Silverman B. 2014. *Modern slavery: an application of multiple systems estimation*. Rep., Home Off., London. https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation

Spiegel PB, Salama P. 2000. War and mortality in Kosovo, 1998–99: an epidemiological testimony. *Lancet* 355:2204–9

Application of multiple systems estimation used by war crimes tribunal to corroborate evidence.

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

Sutherland J, Schwarz CJ. 2005. Multi-list methods using incomplete lists in closed populations. *Biometrics* 61:134–40

Sutherland J, Schwarz CJ, Rivest LP. 2007. Multilist population estimation with incomplete and partial stratification. *Biometrics* 63:910–16

Taylor A, Goldberg D, Emslie J, Wrench J, Gruer L, et al. 1995. Outbreak of HIV infection in a Scottish prison. *BMJ* 310:289–92

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

White SR, Bird SM, Grieve R. 2014. Review of methodological issues in cost-effectiveness analyses relating to injecting drug users, and case-study illustrations. *J. R. Stat. Soc. A* 177:625–42

White SR, Bird SM, Merrall ELC, Hutchinson SJ. 2015. Drugs-related death soon after hospital-discharge among drug treatment clients in Scotland: record linkage, validation and investigation of risk-factors. *PLOS ONE* 10:e0141073

White SR, Muniz-Terrera G, Matthews FE. 2017. Sample size and classification error for Bayesian change-point models with unlabelled sub-groups and incomplete follow-up. *Stat. Methods Med. Res.* In press. **https://doi.org/10.1177/0962280216662298**

Worthington H, McCrea RS, King R, Griffiths RA. 2017. Estimation of population size when capture probability depends on individual states. arXiv:1708.00348[stat.AP]

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