Scotland-wide Data Linkage Framework for Statistics and Research

Sheila M. Bird

Overview, p2: Master-index (initial of first name, soundex of surname, sex, date of birth), as used on Scotland's confidential HIV and Hepatitis C diagnosis registers, made a major contribution to the building of Scotland's reputation and was introduced under MRC-auspices and promoted by MRC Biostatistical Initiative on AIDS/HIV studies in Scotland (MRC-BIAS). Downgrading of master-index to initial, sex, date of birth is insufficient on a number of grounds from wrong order of initials to multiplicity of false matches unless regional data ride to rescue.

Aims of data linkage framework: include b) the production of census-type statistics. Completion of census is a legal obligation. Citizen may opt-out of most other forms of data-collection, to which the legal status of census cannot be transferred. Census inquires about household and/or family structure. Data linkage methods for identifying family-structure are largely untried, and a clearly challenging: females change surname; paternity may be either not acknowledged or unknown; and family-structure evolves.

Aims also include e) increased capacity for robust evaluation of cost-generating events, benefits and risks. Data Linkage Framework should not only increase the statistical power of analyses (such as of serious adverse events in respect of prescribed medicines0 but also the complexity of analyses (for example, to consider co-prescribing or co-morbidities) and sophistication of the questions that can be addressed.

Caveat emptor on e) is that even complex analyses of observational data may fail to adjust fully for systematic biases and there remains an important role for formal experiments which include randomized controlled trials. See also 'what works' on p5.

Section 1 (brief explanation): care should be taken that the framework does not rule out absolutely a potential role for data-linkage in the identification of at-risk individuals. For example, on the basis of matching of master-indices for inmates of Glenochil Prison during January to June 1993 (when HIV transmissions had occurred) against the mater-indices of men who had even been HIV-diagnosed in Scotland, there were distinctly more HIV-infected inmates (potential transmitters) than were identified as such during the infection Control Exercise. Had the data-linkage been done in June 1993 which we did subsequently, the *under-count* would have been known about at the time, and decisions made about how to manage it. The consultation appears to rule out the use of data-linkage in managing outbreaks, which would be a mistake, or in counting HOW MANY appear to be at-risk compared with the numbers so-far-identified by conventional other means, which would be equally regrettable. Thus, do not rule of roles for data linkage either in outbreak management/surveillance or in counting the numbers potentially-at-risk (for comparison with 'known' counts).

Examples (p4): The Health Survey setting is exceptional – in that respondent vouchsafes his/her data and can be invited to give secondary permission for linkage to other data

(such as hospitalisations). Linkage consent-rates should be cited as a matter of good statistical reporting practice. Statistical reporting standards are also honoured in the breach in all three bulleted examples on p5: mean without sd or se; % without denominator etc.

Other national examples could usefully be cited, such as the study by Bird and Hutchinson on drugs-related deaths in the fortnight following release from Scottish prisons in 1996-99, where the contacting of ex-prisoners for 'permission' was impossible, and would have breached the confidentiality of nearly 20,000 - all to find out about 80 deaths from whom permission could not be got, see BIRD SM, Hutchinson SJ. Male drugs-related deaths in the fortnight after release from prison: Scotland, 1996 - 1999. *Addiction* 2003; 98: 185 - 190.

See also a range of papers by McDonald et al. who use record-linkage for surveillance of the late sequelae of hepatitis C virus infection in Scotland, or by Merrall et al. who used record-linkage to analyse cause-specific mortality of clients on the Scottish Drugs Misuse Database and discovered a previously unrecognised, high risk of drugs-related death in the 28 days after hospital-discharge, see for example:

McDonald SA, Hutchinson SJ, BIRD SM, Mills PR, Dillon J, Bloor M, Robertson C, Donaghy M, Hayes P, Graham L, Goldberg DJ. A population-based record linkage study of mortality in hepatitis C diagnosed persons with and without HIV co-infection in Scotland. *Statistical Methods in Medical Research* 2009; 18: 271 – 283.

Merrall ELC, BIRD SM, Hutchinson SJ. Mortality of those who attended drug services in Scotland 1996-2006: record linkage study. *International Journal of Drug Policy* 2011 [29 June 2011 Epub ahead of print].

Benefits (p5): No distinction is drawn between data that I am legally obliged to provide (for example via census or income tax return) versus data that are in my interest to provide (record of my hospitalisation for X, Y and Z) versus data that I vouchsafe (truthful or false completion of health survey) versus medical data that I cannot distort (result of blood test or other investigation from which I do not know the answer when I give permission for the test/investigation to be made). Sometimes, these distinctions matter. In particular, medical data are given under a duty of confidence. Privacy issues and data integrity risks differ – some data I can choose to distort, other data I cannot.

Census-type statistics (p6): For benefits to be realised fully, recourse to capture-recapture methodologies may be necessary; and should be under investigation.

Available to all (p6, last paragraph): some examples of what analyses have revealed might be added.

Investment in data quality (p7): This is an excellent point as shared data are only as good as the data definitions and validated input to the original data-sets, about which – if not previously subject even to logical checks – there needs to be quality-assurance. A particular problem of 'over-writing' can arise with so-called administrative data-sets, whereby information that was previously missing is requested at each subsequent client-

visit, and so presence/absence of the information sought is thus correlated with in-service or survival-time.

There needs to be recognition that investment in data-quality may need to be ongoing as linkage can itself reveal data-deficiencies or data-inconsistencies that a single data-set could not have identified for itself. Also, data-items or questions within a data-set may need to be added to answer well novel questions (which linkage enables) that were not envisaged when the various now-linked data-sets were 'designed'.

Benefit 4 (low cost longitudinal research): This section does not make strongly enough that low-cost applies particularly to the longitudinal record of event-dates (eg hospitalisations, incarcerations, court appearances, tagged-sentences, etc). However, "high-quality" depends on how well-checked event-dates were when recorded on the original data-sets. Errors of recall may mean that my event-history, as I recollect hospitalisations (say), is less accurate for childhood than adulthood; or under-counts repeated spells which were close in time. However, under certain circumstances (for tracking disease transmission when recall of contacts is important), there may be value in knowing the likely patterns of recall-bias, or an extra-covariate for those whose recall of dated-events is particularly good.

P7, footnote reference is incomplete.

P8, missed opportunity: in general, there is need to differentiate short from longer term impacts, such as of widowhood on females' mortality or widower-hood on males' mortality.

P8: before/after policy change is perhaps more often an unnatural experiment than it is a 'natural experiment'.

Benefit 5 (increased capacity): Royal Statistical Society (RSS) has drawn attention to a fundamental aspect of record-linkage that differs – to Scotland's advantage - between Scotland and E&W. Importantly, unlike in E&W, Medical Research Council and RSS note that all deaths in Scotland are registered within 8 days of death having been ascertained, so that - for all persons/patients - survival status can be established *rigorously and in a timely manner* via record-linkage. This does not apply in E&W where completeness of registration is not guaranteed even a year after death.

Not only does Scottish record-linkage allow follow-up to be conducted more completely and economically, it is also achieved without interviewer-intrusion to ask the patient's family about event-dates.

P9, last paragraph: some of the early applications across Scottish jurisdictions might be cited, for example: SEAMAN SR, BRETTLE RP, GORE SM. Mortality from overdose among injecting drug users recently released from prison: database linkage study. *British Medical Journal 1998*; **316**: 426 - 428.

Western Australia, p10: mention should be made of relative population sizes which, in part, determine statistical power (Western Australia: 1 million; Denmark: 5 millions; Scotland: 5 millions; Manitoba: NK millions). Mention should also be made that Western Australia has very high consent rate from its cancer patients (for example) re DNA-provision. Link-in bullet-list omits prescriptions.

The Manitoba example needs to be better described as to whether there is breath alcohol surveillance at antenatal visits as the purpose is presumably to identify the mother's alcohol-dependency and intervene to try to reduce her drinking during pregnancy.

Consultation Q1 – Benefits not sufficiently described include i) public health surveillance where, importantly, a biological sample or test result can be linked to later dated-event sequelae, as for HCV diagnoses. Another benefit insufficiently described is ii) the ability to ask, and answer via more sophisticated analyses, less naïve (or simplistic) questions because of Scotland-wide longitudinal data on dated-events, as should be possible re prescriptions, co-morbidity and serious adverse-events. The third benefit that is not sufficiently described is appreciation of iii) increased analytical competence (as distinct from data-linkage skills, which are also important). Bayesian capture-recapture methods may be needed for estimation of 'hidden totals' and assurance of feed-back loop from analysts to data-collectors so that they understand what their data have revealed and hence the importance of collecting them well.

Thus, on capacity, there are three tasks: task 1 = conduct linkages; task 2 = understand the linked data-set; task 3 = analyse longitudinally-linked data, which is quintessentially statistical but requires subject-matter knowledge to be done optimally. However, tasks 1 and 2 also have statistical aspects and there is ongoing methodological work which weights potential links rather than selects the highest ranked; and in which weights are iteratively updated. Such methodological work can only progress to applications by being tested in safe havens as full access to all potential links is needed.

Consultation Q2: barriers include simplified analyses conducted by persons who lack the professional skills for the complexity that Tasks 2 and 3 may entail: this slows down, and limits, the potential information yield from record-linkage. Perhaps a time-limit on the holding-up of essential analyses could be set? Barriers also include unthinking application of rules which censor access to low but informative counts, even when the potential for deductive disclosure is remote and specific public health importance may attach to low counts.

Principles are generally well-conceived, and well-expressed, except for the section on **consent (see below)**: As a potential problem, I note that an analyst in receipt of linked data-set may in part - on the basis of these very data – complete analyses (as a by-product, say) that, separately, were due for publication as national statistics on a later date. It would be unhelpful if the main research findings could not be disseminated until after the national statistics deadline had passed. UK Statistics Authority has considered this issue and is relaxed about the public good taking precedence – provided, of course, that there was no malign intention to subvert national statistics.

On governance (p15), should information about approved also include the approved research protocol which will have set out analysis plans (much as the protocol for a randomized controlled trial does)?

On governance (p15 & p16, notes 14-16), should there be explicit recognition that there are implicit risks that research-teams run when they surrender some of their access to information explicitly to preserve the confidentiality of individuals. In particular, and rather trivially, the full set of data-checks that I might make had I access to *day* as well as month of birth is not open to me. More seriously, the safeguards that both Seaman et al. (1998) and Bird and Hutchinson (2003) put in place had downsides. A death-in-prison had been missed (because the deceased prisoner's incarceration record was missing as it had been required by his Fatal Accident Inquiry) and we became aware of this only because Dr Brettle recalled that one of his HIV-infected patients had died in Edinburgh prison! Scottish Prison Service (SPS) had provided to Bird and Hutchinson an extra field in which was recorded whether the ex-prisoner was alive on liberation. Thereby, we discovered a 'death' that General Register Office for Scotland (RG)had not notified to us. We assumed that SPS had made a data-entry error, which we brought to their attention. However, there was no error . . . the explanation was that the linkage file that had been passed from SPS to GROS had had spaces typed within it which RGOS'ds program had not coped properly with. The entire linkage was redone but the problem would never have come to light had we not been provided with a data-field we had not even asked for (but clearly should have done).

The time that linked datasets are held must also conform to research-governance.

Consent: Subsections 23 and 20 should be reversed. The present section 20 (on explicit consent) should be re-worded to ensure that those who seek to elicit consent are themselves explicit when they explain to the respondent how data about him/her are held which may be subject to linkage. If I complete a health survey, I may refuse permission for my answers to be subject to linkage if my name is attached to them - even in the safe haven. However, if the health-survey team replaces my name with my master index (S B630 f 180552), then I may be quite willing for linkage to occur. How questions on consent are phrased matters. Public good needs to be balanced against the potentially-biased information loss that consent rates lower than 60% herald.

P18, provision 35: reads as though every clinical trial were obliged to take linkage into account. Some rewording is needed if this, as I assume, is not the intention of para 35.

P19, second last paragraph: to the list of mutual values may be added analytical comparison of the impact of policies which differ in their nature or timing between the nations of the UK, as applies – for example - for take-home naloxone and minimum pricing per unit of alcohol.

P20 Privacy Advisory Service: particular, and different, skills would be needed on part of members of Privacy Access Service for them to be able to make suggestions on i) improved methodology for linkage versus on ii) improved analyses of linked-data-sets.

P22: Caution is needed if methods are to be developed for read-through from, say, master-index for HCV diagnosed patient to patient's identity when the basis for confidential registration was non-nominal. Caution is also need on accreditation of IT systems if, for example, patients would, in effect, be de-registered if their general practitioner or other service-provider was on IT-compatible. UK Statistics Authority's Assessment report on Drug Misuse Statistics Scotland has already picked up on this type of issue as knowing the count of all in receipt of drug treatment in Scotland is more important than whether their provider can supply data by a particular IT system.