Data Capture – for Public Good *(Submission on behalf of Royal Statistical Society [RSS]).*

There is among statisticians generally, in national statistical offices in particular, and in the international statistical literature, considerable experience in issues of data definition and collection **[1, 2, 3, 4]**. This experience deals in particular with the tensions between data completeness and accuracy, and the preservation of privacy.

Disciplines differ, studies differ; funding and time for data acquisition differ, and so does the gamut of questions that any given research study was designed to answer. Principal investigators normally expect to complete, and to publish, their primary analyses before disseminating data more widely – as in the recent studies of bovine tuberculosis (subject to farmer confidentiality).

Rights have complementary duties. The notions of proportionality commensurate with public good and of privacy-rights balanced by citizens' responsibilities to contribute to knowledge are equally important across all fields. Should there be more emphasis on an implied duty to take part in *scientifically and ethically approved* clinical trials, cohort studies, and research-oriented social science surveys; and for principal investigators to set out a time-scale for disseminating data more widely or a process by which others may make applications for access?

Recent controversies have undermined public trust in accredited data capture and their competent management, and put in jeopardy public and professional perceptions about the benefits of research. The balance should be redressed by recalling the sorts of substantive discovery that have been made from data capture for public good. For national databases particularly, we make detailed observations about data quality and the regularity of analyses (see **RSS concerns**).

Broadly speaking, the Royal Statistical Society sees merit in accredited data capture, with analysis, for the public good; and in transparent, *approved* record-linkages either across studies or between databases to create new study-potential. Data-sharing raises pertinent questions **[1]** about ownership, consent for data-sharing, scientific purpose and methods, and permissions for data release: when, why, to whom, collaboratively or competitively, and under what safeguards. There are technical issues to be resolved, particularly in respect of record-linkage, as set out in RSS's **Points of Note**. Scientific standards need to be met by those who create new study-potential by data-sharing. We caution that those who collected data may have considerable 'tacit knowledge' that may not have been fully documented.

The Royal Statistical Society makes **three recommendations**, endorses the *Rawlins principles*, and puts forward 15 **Points of Note** (**Section 3**). **The RSS recommendations are:**

a) Standards of data management need to be sufficiently high that research data **can be** shared for the public good – such as to create new discovery-potential;
b) For transparency, national databases should have a publicly-available protocol which describes the data held, their regular analysis, and any approved record-linkages; and
c) Better public understanding is needed about databases, their linkage, and value-added analyses. The Royal Statistical Society's Getstats campaign could contribute to this goal.

Rawlins 1: Safeguard the well-being of research participants.
Rawlins 2: Facilitate high-quality research to the public benefit.
Rawlins 3: Be proportionate, efficient and co-ordinated.
Rawlins 4: Maintain and build confidence in the conduct and value of research through independence, transparency, accountability and consistency.

1. Background

Broadly speaking, the Royal Statistical Society (RSS) sees merit in accredited data capture, with analysis, for the public good; and in transparent, approved record-linkages either across studies or between databases to create new study-potential. The ESRC gave an early lead. Its Research Data Policy, which required all research grant award holders to offer data collected during the course of their research for preservation and sharing, has recently been updated [1].

In 2011, a short-life RSS Working Party on Data Capture – for Public Good is to report on statistical principles and practice that support data capture for the public good; and to recommend developments in probabilistic linkage and analysis to safeguard the public good in data-sharing.

The RSS Working Party builds on substantial progress made by others in three key reports [2 3 4]. The *Data Sharing Review* in 2008 [2] addressed **why** is it appropriate to share personal data for a particular purpose (answer – because proportionate) and **how** (which data are to be shared, and by what means). In *Sharing research data to improve public health* [5], funders of health research endorsed the **how**: entrench standards of data management so that research data can be re-used effectively; professional recognition for data-management; and due acknowledgement by secondary analysts to data-generators.

However, in 2009, Anderson et al. [3] had called into question the legality, effectiveness and cost of the *Database State*. The UK Government has built, or extended, central databases that hold information from health and education to welfare, law-enforcement and tax with the intention to make public services better or cheaper; but has been challenged by controversies over effectiveness (NHS Organ Donor Register), privacy (Revenue and Customs), legality (National DNA Database) and cost (NHS Detailed Care Record). Many question the consequences of giving increasing numbers of civil servants, and others, daily access to our personal information.

In spring 2010, the UK Government invited the Academy of Medical Sciences to review the regulation and governance of health research involving human participants, their tissue or their data. In January 2011, the Academy's working party proposed: *A new pathway for the regulation and governance of health research* [4] to resolve the delays, complexity and inconsistency across the regulation pathway; to address a lack of proportionality in regulating clinical trials, and inappropriate constraints on access to patient data; and to bring about a cultural change in healthcare to promote, and value fully, the benefits of health research.

The dual notions of proportionality commensurate with public good and privacy-rights balanced by citizens' responsibilities to contribute to knowledge are equally important outside of healthcare. **Medical data are different**: in particular, biological samples for diagnostic and other testing are obtained by doctors under a strong duty of confidentiality, and for declared purposes. The duty of confidentiality is crucial because test results **may reveal information hitherto unknown, even to the patient,** and which the patient cannot rescind without recourse to falsification of, or deletion from, their medical record. Neither action is in the interest of either the patient or epidemiology.

Recent controversies have undermined public trust in accredited data capture [6 7 8 9 10] and put in jeopardy public and professional perceptions about the benefits of research. We redress the balance by recalling the sorts of important discovery made from data capture for public good [11 12 13 14].

2. Royal Statistical Society's concerns and recommendations

Particular concerns are:

i) **Delayed discoveries** (whether about benefits or harms) for want of insightful analysis of healthcare and other databases;

ii) **Limited scope for discovery,** because fundamental statistical issues of data-definition and recording have been overlooked in the design of national databases [7, 8, 15];

iii) **Databases whose potential is marred** because records are subject neither to logical checks nor to timely formal analyses which risks persistence of data-errors which mar their potential;

iv) **Improved public understanding** of databases, their linkage, and value-added analyses, a goal for Getstats.

A range of data-quality issues comes to light **only** when data are checked for logical inconsistencies not only within an individual record but across serial records - either within or between databases. A second set of data-quality issues may be revealed when logically-checked data are subject to formal analyses using multi-factorial methods, and inspection of "goodness of fit" shows up outliers or other surprising features in posterior distributions.

RSS recommendations:

a) *Standards of data management need to be sufficiently high that research data* ***can be*** *shared for the public good – such as to create new discovery-potential;*
b) *For transparency, national databases should have a publicly-available protocol which describes the data held, their regular analysis, and any approved record-linkages; and*
c) *Better public understanding is needed about databases, their linkage, and value-added analyses. The Royal Statistical Society's Getstats campaign could contribute to this goal.*

3. Points of Note by the Royal Statistical Society

N1. Journals limit word count in print-editions. Scientists may prune methodological detail to such an extent that others lack important detail for reproducing their method, or computing code. *Web-appendices are a boon: for methodology and computing code to be explained in detail.*

N2. In statistical science, investigators may analyse the same data by another technique which makes a different set of assumptions. Inferences may turn out to be broadly similar, so that re-investigation has been primarily of academic interest; or may be radically different with important implications for public policy. *Insightful re-analysis generally requires access to the original data, rather than to summaries thereof.* Some journals require, rather than invite, authors to deposit their data-sets on the web as a condition of publication but an exceptional case can be made on commercial or other grounds, for example potential harm to individuals or risk to intellectual property.

N3. Permission to access already-analysed non-nominal data – for example, capture-recapture counts giving the numbers of injecting drugs users whose identifier featured on one or more of four data-sources (A, B, C, D) per combination sex, age-group and region - can be time-consuming. Publication of the received counts may be barred as a pre-requisite for access. The statistician-scientist has to choose between: inability to discover if/that different methodology gives an importantly different answer; and, contrary to good scientific practice, to accept an initial bar on publication of the received-counts in the hope that an importantly different answer will ensure later-lifting. *Data-owners, including governments, place restrictions on data-recipients, some of which have an impact on the potential for peer-review publication.*

N4. Even such an apparently simple data-request as the above cross-counts can create substantial, and usually unfunded, work for the data-holder if, for example, the requested age-groups (under

25, 25-34, 35+ years) were different from those analysed originally (say: under 35, 35+ years). Freedom of Information Act recognises that applicant-incurred costs may need to be capped. *There needs to be recognition of, and re-imbursement for, data-management costs that may be entailed in meeting the data-sharing requests of bona fides scientists, let alone of 'citizen scientists'.*

N5. **When to share, which data, and why:** Scientists exercise professional judgement in deciding: a) which data-sets or databases to request access to, b) which specific data-fields within a database are pertinent to their scientific inquiry, c) how to support their request by providing an outline protocol for the investigation which uses the requested data, and d) how they would propose to collaborate with, or acknowledge, the data-provider. The same standards should apply to 'citizen scientists'. *Citizen scientists may require resourced professional support to meet the required scientific standards.*

N6. *Perusal of study-design or investigative-protocol is often sufficient to persuade a professional scientist that a study's data would not be worthwhile acquiring.* For example, the experimental-design may have been inadequate for comparative inference, see Blueprint [16].

N7**. Central to any scientific enterprise is insight:** *Knowing which questions to ask, and how to pose them – together with any qualifiers - ensures that acquired data can be efficiently-analysed.*

N8. *Opening up of scientific information should not be at the expense of the conventional scientific standards (peer-review, competitive-funding, ethical-review, data-security) that a research-participants rightly expected when consent was originally given (in a Medical Research Council study, say).* Warnings there are a-plenty that consent-rates generally have declined - in national surveys, to cadaveric solid organ donation (down from 70% (se 1%) in 1990 to 60% (se 1%) in 21st century), in Biobank (well below 40%).

N9. **Web-based surveys need further improvement:** Recipients to whom a web-based survey has been emailed should be told: i) exactly how, and by whom, their survey-answers and email-address will be unlinked; ii) whether any information from their email address (such as "ac" for academic or "cam" for Cambridge or "mrc" for Medical Research Council) will be retained with their survey-answers; iii) how the list of email-recipients was sourced; and iv) when, and where, the analysis of survey-answers will be obtainable. Finally, web-surveys should be so-designed that v) a potential-respondent can read through, or print-off, the full set of survey questions ***before*** deciding whether to take part. Web-based surveys which breach these requirements are a disservice to citizen-scientists' understanding of how scientific method ought to be explained to potential-participant, and deployed to prevent deductive disclosure about respondents. *Improving the credentials of web-based surveys is a task for RSS's GETSTATS campaign.*

N10. **Scientific method has nothing to fear from openness:** Investigative-protocols – or study designs - should already have been written down in sufficient detail to ensure that the study-plan can be implemented rigorously, with any protocol-amendments logged by the date of their implementation. *With few exceptions, subjects should have right of access to the protocol for any study for which their informed consent is requested*. Some psychological experiments are exceptional, see [17].

N11. **Learning from controversies:** The discovery-potential from analysing existing non-research data can be seriously compromised if it transpires that data have accumulated over many years, such as for administrative purposes, without having been subject either to logical checks or to much in the way of formal analysis. *Minimal checking risks minimal data-quality, and may*

compromise analyses that, in principle, were do-able but in practice are not. For example, the Police National Computer (PNC) records sentence date and the sentence(s) handed down on that date, but does not reliably record whether sentences run concurrently or consecutively and so, in practice, PNC records do not identify when offenders are incarcerated – as this was not a question that PNC was designed to answer! Likewise, for more than 10 years, the Scottish Drug Misuse Database recorded the start of new treatment episodes and drug-behaviours then, but not the end-date for any treatment-episode and so could not be used to answer questions about prevalent, and as well as incident, clients. Such questions are important nonetheless. A third example is the NHS Organ Donor Register on which there were over 15 million registrations before it was discovered that, for about 10 years, organ-specific permissions by those who had registered their willingness to donate via DLVA had been transposed and this resulted in a loss of cadaveric heart donations [9]. *The RSS suggests that no national database should be held for more than 3 years without formal checks on data-quality and technical review to ensure that the data held are fit-for-analytical-purposes: in terms of the major questions that analysis of that database should be capable of answering. In short, there is a requirement to analyse, both to discover and correct lacunae in the data-held and in data-quality.* The National DNA Database was found wanting in terms of data-held when the Home Office sought statistical defences for how long to retain the DNA-profiles of the "innocent".

N12. **Record-linkage:** Scientists engaged in secondary data-analysis recognise that a level of data-checking that would have been available to them as data-gatherer is not, namely: look-back to source-records to verify data-abstraction. *The secondary-analyst is reliant on the standards that the data-gatherer applied. Scientists who create discovery-potential by record-linkage (without explicit consent) across a number of databases face additional limitations . . .* The investigative-scientist may be the data-holder for database A, one of three linkable databases (A, B, C) and, by linkage, could learn about the B-history and C-behaviours of all A-registered persons. The need for 'safe havens' for record-linkage becomes immediately apparent as A-participants (for whom the investigative-scientist may hold identifying details) did not give informed consent for disclosure of their now-linked-in, and hence identifiable, B&C records. However, if the investigative-scientist either analyses the linked-records **only in a 'safe haven'** (where reference to A-participants' identities is impossible) **or determines the analysis plan that 'safe haven' analysts implement**, then deductive disclosure about individual A-participants is obviated. Non-nominal, without consent record-linkage typically uses minimally-sufficient identifiers (S B630 f 180552). Probabilistic record-linkage on minimally-sufficient identifiers inevitably makes some wrong linkages, whereby individual I's A-record is linked to individual J's B-record because I and J share the same minimally-sufficient identifier. There is a series of technical issues about how to conduct probabilistic record-linkage: some of the more sophisticated methods can only be put into practice in a 'safe haven' because they require simultaneous access to the master databases (A, B, C) rather than to the subset of the B and C records that appear to be 'best-linked' to an A-record. Indeed, depending upon the proposed analysis, and the degree of logical checking across potentially-linked records, the definition of 'best-linked' B-record will itself change [18].

N13. **Ascertainment bias:** Inference - based on potential discoveries about the B-history and C-behaviours of the A-registered subjects above – may be restricted by how the A-register was itself compiled. Diagnosis-registers may be subject to ascertainment bias, if diagnosis of a blood-borne virus, such hepatitis C virus carriage, tends to be made later in the patient's incubation period or at post-mortem. *Analysts need to consider carefully how ascertainment biases pertaining to each linked register may affect their ability to generalize findings from a record-linkage study.*

N14. **Incentives:** Shoppers receive modest financial incentives, calibrated to their monthly spend, to encourage their signing-up to supermarkets' databases. *Those who sign-up, in effect, allow stores to relate the payee's demography, family size and income bracket to their purchases in the premises of one or more store-chains.* The modest financial incentives that suffice to recruit customers, who account for **x%** [not known] of a store's revenue, to have their various purchases linked - both over time and to their sign-up details - suggest that the public is not overly-precious about disclosure nor about commercially-detailed analysis of its nutrition, drinking pattern, travel locations and pharmacy purchases. *The RSS notes also public's tolerance of social networking sites behind which lie data-mining techniques used for targeted advertising.*

N15. **Intellectual property, elephant in the room:** Too often still, key public policy decisions are informed by evidence that is made available early to decision-makers but which is not in the public domain until accepted for peer-review publication, which may be long delayed. *The NICE model, of publicly-funded assessment reports which are made available to consultees during the appraisal process, has major merit.*

4. Answers to Royal Society's questions.

Q1. *Governing ethical and legal principles.* Please see RSS recommendations. Please see RSS Notes N5, N8, N10 and N15.

Q2a *Application of principles to publicly-funded research.* Please see RSS concerns i) – iv) and RSS recommendations. Please see RSS Notes N4, N5, N8, N10, N12 and N15.

Q2b *Application of principles to privately-funded research on/about individuals or organisations.* As for Q2a. Please also see RSS Note N14.

Q2c *Application of principles to research that is entirely privately-funded but will possible public implications.* Please see RSS concerns i) and ii), and RSS recommendations a) and c).

Q2d *Application to research/data sharing that involves the promotion of the public interest but could have implications for the privacy interests of citizens.* Please see RSS concerns i) and ii), and RSS recommendations. Please see RSS Notes N1, N3, N9, N11, and N12.

Q3. *What activities could improve the sharing and communication of scientific information?* Please see RSS recommendations. Please see RSS Notes N1, N2, N3, N4, N5, N7, N8, N11, N12 and N15.

Q4. *How should new media change how scientists conduct and communicate research?* Please see RSS recommendation b) and RSS Notes N1, N5, N6, N9 and N14.

Q5. *Additional challenges in making data usable by scientists in the same field, other fields, or by 'citizen scientists'?* Please see RSS recommendation c) & Notes N1, N3, N4, N5, N6, N10, N15.

Q6. *Benefits of more widespread sharing of data?*
Please see RSS concerns i) to iv) which are balanced by RSS Notes N5, N6, N7, N12 and N13.

Q7. *How should concerns about privacy, security and intellectual property be balanced against openness?* Please see RSS Notes N4, N8, N14 and N15.

Q8. *What should be expected of scientists/others?* Of scientists, that they adhere to professional codes and act in the public interest; of others, that they respect similar standards.

References

1. Keiding N. (with Commentary by Breslow NE; Cox DR and Donnelly C; Deangelis CD and Fontanarosa PB; Donoho DL; Goodman SN; Groves T; Peng RD). Reproducible research and the substantive content (with Commentaries). Biostatistics 2010; 11(3): 376 – 396.

2. Official Statistics: Counting with Confidence. Report of a Working Party on Official Statistics in the UK (Chair: Professor Peter G Moore). *Journal of the Royal Statistical Society Series A* 1991; 154: 23-44.

3. Royal Statistical Society Working Party on Performance Monitoring in the Public Services (chair: Professor Sheila M. Bird). Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society Series A 2005;* 168(1): 1 – 27.

4. Royal Statistical Society Working Party (chair: Professor Stephen Senn). Statistical issues in first-in-man studies. *Journal of the Royal Statistical Society Series A (Statistics in Society*) 2007; 170: 517 – 579.

1. ESRC research data policy changes; are you (and your data) prepared? http://www.esds.ac.uk/news/newsdetail.asp?id=2884
2. Thomas R, Walport M. *Data Sharing Review*. Information Commissioner's Office, London: July 2008. (ISBN 978-1-84099-204-5).
3. Anderson R, Brown I, Dowty T, Inglesant P, Heath W, Sasse A. *Database State.* Joseph Rowntree Reform Trust, York: March 2009. (ISBN 978-0-9548902-4-7).
4. Working Party of the Academy of Medical Sciences (chair: Professor Sir Michael Rawlins). *A new pathway for the regulation and governance of health research.* The Academy of Medical Sciences, London: January 2011. (ISBN No: 978 -1- 903401-31-6).
5. Walport M, Brest P. *Sharing research data to improve public health. Lancet* 2011; 377: 537 - 539. Published online January 7, 2011 as DOI:10.1016/S0140-6736(10)62234-9.
6. Bird SM. Real root of a data disaster. Sunday Herald 2007; 2 December : 30.
7. National Police Improvement Agency. National DNA Database Annual Report 2007-09. National Police Improvement Agency, London: mm 2009.
8. Hawkes N. Home Office keeps the lid on DNA-retention analysis. *Straight Statistics* 2011; 12 July. (see http://www.straightstatistics.org/article/home-office-keeps-lid-dna-retention-analysis).
9. *Review of the Organ Donor Register by Professor Sir Gordon Duff.* DH, London: 19 Oct. 2010. (http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_120563).
10. Press briefing: 2011 Census and Lockheed Martin UK .*https://2011mc.census.gov.uk/index.php?module=documents&action=view&id=669. See also http://www.guardian.co.uk/uk/2011/feb/19/census-boycott-lockheed-martin.*
11. Hutchinson SJ, BIRD SM, Goldberg DJ. Modelling the current and future disease burden of Hepatitis C among injecting drug users in Scotland. *Hepatology* 2005; 42: 711 – 723.
12. Drug Misuse Declared – Findings from the 2010-11 British Crime Survey: England and Wales (edited by Smith K and Flatley J). Home Office Statistical Bulletin 12/11. (see http://www.cjp.org.uk/publications/government/drug-misuse-declared-findings-from-the-2010-11-british-crime-survey-england-and-wales-28-07-2011/).
13. Generation Scotland. See http://www.generationscotland.org/index.php?option=com_content&view=article&id=46&Itemid=28#H1N1
14. Cognitive Function and Ageing Studies. See http://www.cfas.ac.uk/.
15. 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: accepted for publication.
16. Goldacre B. Blueprint fail. *Guardian* 2009; 19 September. See also http://www.badscience.net/2009/09/blueprint-fail/
17. Senn SJ. Are placebo run-ins justified? *British Medical Journal* 1997; 314: 1191-1193.
18. Goldstein H, Harron K, Wade. The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine* (under review).