

Patient Record Data: Statistical Disclosure Control for Grid Based Data Access

M. Elliot¹, K. Purdam¹, D. Smith¹

¹Cathie Marsh Centre, University of Manchester

Mark.J.Elliot@manchester.ac.uk

Abstract. Patient record data is potentially highly sensitive and its secondary use raises both ethical and data protection issues. Disclosure of patient data could cause serious difficulties for the medical profession and be potentially damaging for individual patients and clinicians. Yet at the same time patient records are a hugely valuable resource in terms of clinical research and patient treatment. A secure, remote access system for such data would provide numerous benefits.

In this paper we outline the statistical disclosure risks posed by patient record data in the context of the establishment of a grid based medical data repository. We review good practice in existing patient databases, outline a scenario model for assessing risk and suggest a new model for statistical disclosure control of patient data. The findings are likely not only to be of interest to health researchers and practitioners, but also to serve as an exemplar to the development of e-research for grid developers and users across different data types and policy areas.

Introduction

Under the Data Protection Act 1998 patient record data is considered “sensitive personal data”. The Data Protection Act 1998, the Human Rights Act 1998, and the common law duty of confidence all protect the privacy of patients and their medical information.

It is commonly asserted that health information that identifies individual patients must not be used without the patients’ informed consent. Data that does not identify a patient (under the Data Protection Act definition) does not require consent. In addition, the Data Protection Act allows exemption for ‘medical purposes’ which includes medical research; Section 39 allowing for statistical or historical research.

This leads to the question of how we assess whether or not a given set of data allows the identification of, or disclosure of information about, a patient. Here we enter the complex arena of statistical disclosure risk assessment. Statistical disclosure, has been defined as (Elliot, 2005),

“the revealing of information about a population unit through the statistical matching of information already known to the revealing agent (or data intruder) with other anonymised information (or target dataset) to which the intruder has access either legitimately or otherwise”.

This definition underlines the point that to protect confidentiality, it is not sufficient to simply remove what are called direct or formal identifiers (such name, address, and NHS number), as disclosure could happen through statistical matching of other information. A recent example is the identification of a late abortion case (both clinician and patient) through the release of tabular statistics with small cell counts.

The focus of this paper is on statistical disclosure control and the risks to individuals posed by the determined intruder with access to tabular or analytical outputs of patient record data. The research reported here looks specifically at the development of patient record repositories such as the one being developed under the Clinical E-Science Framework (CLEF).

Context

In health care it has been argued that the protection of privacy in medical records has inhibited service provision and research. Cancer Research UK has claimed that the requirement, under the Data Protection Act, to protect confidentiality through anonymisation is having a detrimental effect on research (see *The Observer* 5.10.03). The charity is campaigning for an exemption from the Data Protection Act for medical research. The rationalisation of databases in health and social care, may lead to improved continuity in patient assessment, improved treatment, avoidance of adverse drug reactions, and ensuring that people at risk (including children) receive appropriate support services. Confidentiality requirements may have a damaging impact on the tracking of communicable diseases, outbreaks, vaccine effectiveness and failure to recognise new strains. Conversely, without such consent procedures individuals may be unwilling to provide such information, which would also weaken the information resource. These concerns can be traced back to the Caldicott Report (1997), which highlighted the need for the ongoing review of all patient-identifiable information.

Recent research by the NHS (2002) has highlighted that there are high levels of trust amongst the public about the protection of patient confidentiality, but only limited knowledge about how their information is used. The research suggests that the public are uncertain who has access to their records and under what conditions. When asked if they would want to remove particularly sensitive information from their shared electronic record, 60% of respondents said that they would remove nothing, 24% said they would remove a little, and 4% each said they would remove "a lot" or "all". The research concludes that for the majority of people there is little or nothing that they consider sensitive on their health record. However, they would still like to have some control over what happens to their information. Regarding data sharing, the respondents felt that the NHS should strive to achieve anonymity for all patients if any information was to leave the NHS system. In addition, the Alzheimer's Society in the UK has argued that not only can it be questioned whether all research can be justified in terms of the public interest, but there is a need to involve patients in the development of any new access framework for the use of patient data (Cayton and Denegri 2003).

It is notable that the ONS is presently consulting on good practice in the publication of tabular health data (see ONS 2006). The consultation was prompted by concerns following the identification of a patient undergoing a late abortion and of the clinician treating her. See Kluge (2001) for a discussion of the ethics of electronic patient records.

The release of census data provides a useful comparison when considering the tension between data protection and data use. The 2001 UK Census data is only released in tabular form and is subject to a range of disclosure control methods. These have an impact on the

quality of the data (Purdam and Elliot, forthcoming). Such techniques include reducing sample sizes, perturbing, rounding, swapping and adding noise (see Marsh, 1991 and Doyle *et al.*, 2001). This is despite the fact that the range of information is much more limited, and arguably much less sensitive, than that which is available elsewhere. Arguably the census provides a useful exemplar for the development of access to medical data and the need to make data utility a priority. There is little point putting resources into making patient data available if it is not released in a form that has enough detail to make it useful, or even worse has been perturbed to the extent that it does not have the required level of accuracy (Purdam, 2006).

In relation to medical and patient record data, potential data intruders might include journalists, an unhappy patient, someone with knowledge of a patient, insurance companies, pharmaceutical companies, political organisations or medical professionals. As outlined in recent research by the NHS (2002), if health information is inappropriately shared outside the NHS it may prejudice people's ability to get jobs, life insurance or mortgages. Information shared inappropriately within the NHS could affect the way people are treated by health and other public services.

The consequences of a disclosure of patient record data may be far reaching, not only for the individual patients and / or clinicians, but also for the future access to valuable patient record data. Even so, within the NHS and in medical research more generally, accidental or deliberate disclosure of medical data is not unprecedented. Disclosure has taken place through the theft of computer equipment, the holding of data on unsecured servers, the discarding of computer equipment without the deletion of relevant files, and data being sent erroneously as e-mail attachments. There are also other examples of disclosure: an image of a skin complaint, which had been used without the patient's permission in the British Medical Journal, was recognised by the patient. In relation to published health statistics, a clinician and patient were identified from a published table of abortions statistics. The identification was made from the low number of late abortions given in the table and the region where the patient lived and the clinician worked. For a summary of recent disclosures see Rogers (2005).

Background to CLEF

CLEF is focused on developing methods for managing and using pseudonymised repositories of long term patient histories which can be linked with genetic, genomic, and image information or used to support patient care. One of the main aims of CLEF is to establish a pseudonymised repository of histories of cancer patients. The main users of CLEF are to be clinicians and scientists. Clinicians will be able to request detailed patient records over a person's whole life. The aim is that clinical researchers will be able to select particular samples of patients by condition, symptoms, treatment and by particular demographics or genomic information.

All information passed to the CLEF repository by the data provider (an NHS organisation or trusted third-party) will have been pseudonymised by the removal of all obvious identifiers such as name and date of birth. A CLEF entry identifier will have been attached to each case. This identifier can only be used to re-identify the patient by the data provider. Any residual identifying information such as names of relatives, nicknames, and names of GP or clinicians will also be removed within CLEF before being passed to the research repository. The information is then made available to clinicians and researchers in a range of predetermined

tabular outputs. Disclosure control proposals for CLEF have included cell size restrictions and the blurring of results (see Kalra *et al.* 2003).

The approach of CLEF has been to assume that if individual records can be read in detail, there is a risk of identification, and so all information should be treated as if it were identifiable (Taweel *et al.*, 2003; Kalra *et al.*, 2003). However, the view is that the actual risk of identification is low. For access to data or subsets of data beyond predetermined outputs or metadata, permission is required from the Ethical Oversight Committee. Again data access will be closely monitored. Moreover, access to a full patient record is not likely to be given. A range of parallel CLEF projects are developing privacy enhancing technologies including secure grid infrastructure, data encryption and access protocols. See <http://www.clinical-science.org>.

A unique feature of the CLEF repository will be that it will be possible to link back the original patient ID in order to contact high risk patients or to recruit patients to trials. This will only be done through the originating hospital who will take responsibility for contacting the patient. Researchers will have to pass pseudonyms back to CLEF, which will be re-translated to the originator identifier (another pseudonym). This will be passed back to the originator and linked back to the original identifier (NHS Number or Hospital Number) so that the patient can be contacted via the treating clinician.

Below we summarise the outcomes of our consultation with existing patient and medical record database managers, give an assessment of the availability and risk implications of medical data in the public domain, and the outline our new model for the statistical disclosure control of patient record data.

Methodology

This paper brings together findings from qualitative policy research, risk assessment and analysis of disclosure risk. The policy and practices review involved consultation with managers of existing patient record databases in order to examine their approach to disclosure control. The risk assessment involved the development of statistical disclosure attack scenarios. This methodology was developed by Marsh *et al.* (1991) and Elliot and Dale (1999). Elliot and Dale's classification scheme takes account of the data intruder's perspectives. The 11 point scheme is as follows: motivation, means, opportunity, attack type, key matching variables, target variables, effect of data divergence, likelihood of success, goals achievable by other means, consequences of attempt and likelihood of attempt. Our initial work in this area has focused on (i) assessing the amount of health information in the public domain, by analysing the information collected by health organisations; (ii) by conducting intruder simulation experiments where specific patient level information is sought in a fixed research time.

The disclosure risk assessment method is similar to that of Karr *et al.* (2002) whereby the risk is assessed in terms of the bounds that can be placed on the cell counts in a 'base' table given a release of some of the base table's smaller margins. It differs in a number of respects; mainly in that releases are not considered to be released to 'the world'.

Findings

Consultation with patient record and health database managers

We interviewed and reviewed the approaches to disclosure control of five gold standard patient record and medical data sets.

Our review highlighted that even amongst medical databases that are thought to have gold standard data handling practices there is much variation in confidentiality and disclosure control practices. Whilst all such datasets have direct identifiers removed during the data preparation process, and there is recognition of the risks associated with geographical information, similar data are protected to different degrees across different datasets. Not all data providers track user requests between database updates. We also discovered that there are varying approaches to risk. One database provider took a pragmatic approach to confidentiality, specifically to the presence of unique individuals in certain databases. The complexity of certain medical information means that rare cases and combination of cases will often be identifiable. Yet at the same time such cases can often be the focus of research interest.

Different rounding and banding policies are used but the precision required for reliable statistical inference limits the extent to which data can be perturbed. Some, but not all databases do systematic analysis of disclosure risks from unique cases. Measures can also vary for data from different countries to reflect data content, disclosure risk assessment and data protection regimes. There is only limited systematic analysis of the potential for disclosure from published analyses.

Availability of individual health data in the public domain

We also assessed the increasingly availability of individual health data in the public domain. This Data Environment Analysis (DEA) methodology provides a detailed assessment of the type of information collected across service providers and data held by commercial data providers. An accurate measurement of the availability of data in the data environment of a given dissemination process will ensure that any disclosure control techniques are as efficient as possible, limiting the loss of data utility.

DEA provides accurate, up to date information on individual variable availability. This can be used to drive the construction of scenario frames as part of the disclosure risk assessment process. CLEF will need to take account of the availability of individual data and possibly disclosive data from other sources.

In the first phase of our study, over a two week period we analysed fifty four paper and electronic forms used by a range health organisations and online discussion groups. Specifically we considered: insurance companies, job application forms, health club membership, online health forums and discussion boards.

From the sample of forms a total of 1168 variables could be generated. Many of these duplicate the type of demographic information that might be found in an anonymised data repository. Information included: demographics, GP name, symptoms, condition, medical history, diagnosis, treatment, whether part of a clinical study, partner details and leisure details. More specific examples from online cancer discussion forums included first names, clinician names, treatment, drugs and drug reactions.

The meta-dataset that we have created of these variables is available via an HTML interface. The interface allows searching by variable type or the whole meta-dataset can be downloaded.

This illustrative DEA gives an initial snapshot of the scale and type of information gathering, and the data held in restricted access datasets and in the public domain. This data forms part of the wider data environment in which e.g. the CLEF database will have to operate. It is important to assess how the availability of information and type of information changes over time. We will extend this work by looking at the availability of the some of the specific variables proposed for inclusion in the CLEF database. Further risks could be posed by a more determined intruder who engaged in direct contact with the users of online discussion forums. It is likely that much more detailed information could be collected.

Attack resourcing simulation

Attack resourcing simulation is a research methodology based on giving a third party a specific information gathering task. It allows an assessment of what information is in the public domain and the time taken to find such information. As such, it gives us an insight into the level of time, effort and resources required by an intruder who is gathering identification information from public available sources.

In this case, an independent social researcher was asked to find as much information as possible about a specific hospital's patients within an eight hour period. The simulated intruder was specifically asked to find out any information about any patients being treated by the hospital. Using a range of search strings and different internet search engines the researcher looked for patient lists, names, conditions, treatments and clinicians.

In the allowed time over seventy sources of information were identified. These included: newsletters, news archives, press releases, charity campaign materials, patient diaries, clinical trial reports and research project reports, and accounts of treatments and patients names. Specific information included: patient's name, age, occupation, condition, treatment, relatives' names, region, patient's photograph and clinician's name.

A more extensive research agenda could be developed in this area, combining the data environment assessment and attack resource simulation with web crawling technology. Automatic simulated attacks could be run using synthetic databases generated from real datasets in the release environment. Such an approach would avoid the need for bootstrapping risk analyses based around static key variable definitions. This kind of system could be running continuously, and could alert a data disseminator as soon as there was a potential threat in the environment.

It is also worth noting that such tools are potentially useful to a data intruder and, although not plausible at present, it may soon be necessary to conduct experiments to generate a web harvesting class of scenarios.

Disclosure risk assessment

Definitions of statistical disclosure generally involve one or both of the following:

Identification – A one to one association between a data unit and a target.

Attribution – The association of one or more variable values with a target.

A data unit is an individual or organisation contained in microdata or tabulated data. A target is a data unit about which a data intruder is trying to discover information.

In some cases it is possible for an intruder to perform identification or attribution with absolute certainty (subject to obvious assumptions regarding the correctness of the data). In these cases the identification or attribution is termed exact. Otherwise the identification or attribution is termed approximate.

Exact identification is possible only if a population contains unique records; and exact attribution is possible only if a population cross-classification contains zeros (Smith and Elliot, 2003).

Example:

Table I is reproduced from a U.S. Department of Commerce report (1978). Conditioning on a target being a resident of County B implies that the target is black. The report explains that a risk of such exact disclosure exists if a marginal total (in dimension n-1) of the population table equals one of its detail cells (in dimension n). Clearly this implies the presence of zeros on the base table. But even when there are no marginal totals equal to a detail cell, the presence of a zero allows the intruder to infer that the corresponding combination of values does not apply to the target. (This possibility is discussed in the U.S. Department of Commerce report, but the authors do not consider that such inferences constitute disclosure.)

	Race			
County	White	Black	Other	Total
A	15	20	5	40
B	0	30	0	30

Table I. Number of beneficiaries by count and race.

Conversely we might consider the case when all the detail cells are non-zero. This is the only situation when, regardless of the target and the information held on the target by the intruder, any possible combination of unknown values remains possible. There are a couple of important points to note. Firstly, we should not be concerned about structural zeros, as these can be inferred without reference to the population table and, by definition, correspond to impossible combinations of values. Secondly, the inferences described above do not depend on the presence of uniques in the data; they only depend on the additional knowledge that a target is a member of the relevant population, which in turn is implied by the term ‘exact’.

A further consideration for exact attribution (and identification) is the possibility that the intruder has knowledge of other members of the population and uses this to facilitate disclosure. The most obvious example is where the intruder is a member of the population and can remove his / her own record from the dataset, producing a residual population table that contains a zero (or one) which was not present before. The U.S. Department of Commerce report discusses the more general problem of ‘coalitions’ of individuals within a data set who might cooperate in order to discover new information about targeted individuals. Risk measures that take into account additional knowledge of the population are discussed in Smith and Elliot (2003) and Smith and Elliot (2006).

The above discussion suffices to demonstrate that for exact disclosure it is the low counts in population tables that are particularly problematic, and this most definitely includes counts of zero. This is not necessarily the case for approximate disclosure. For discussions relating to approximate disclosure see, for example, Skinner (1992) and Duncan and Lambert (1986).

Data release issues

A correct risk assessment must take account of the total information released to a user, rather than just the current data request. Typically a user will request a number of tabular outputs from an online data repository over a period of time. If we restrict discussion to the release of marginal population cross-classifications, then the marginal counts contained in the released tables represent a set of linear constraints on the counts in the base cross-classification over all the variables in the released tables. The lower and upper bounds that an intruder can place on the counts in the base table can be used as the basis of measures of risk. However, solving the bounds for arbitrary sets of released tables is computationally demanding, and infeasible in most practical circumstances. Also, it is easy to construct examples where marginal tables with arbitrarily high counts imply zeros in the base table. But there are classes of release for which the bounds calculations can be performed efficiently. The most important class is those releases that correspond to decomposable graphical models (Dobra and Fienberg, 2000). Details of the bounds calculations can be found in Dobra and Fienberg's paper. General information of graphical models can be found in Lauritzen (1996).

Data utility

Many disclosure control methods rely on adding noise to the exact data by, for instance, changing counts in cross-tabulations. Whilst this often (but not always) prevents an intruder from making reliable inferences about targets, the results of subsequent statistical analyses can also be affected. Thus, there is an incentive to release exact data when possible. Alternatively, data might be suppressed, rather than perturbed. Suppression might involve obscuring individual counts in cross-tabulations, or refusing to release certain cross-tabulations altogether. Choosing the most appropriate method for a given application is not generally easy.

One problem with managing disclosure risk by the suppression of data is that 'high utility' data might have to be suppressed because of a previous release of 'low utility' data. There may be many maximal data releases that satisfy a given risk criterion, but they might have varying utilities.

Karr *et al.* (2002) consider the enormous computational benefit of restricting releases (of exact cross-classifications) to those that have decomposable graphical representations, and consider the problems of restricting high utility releases because of previous low utility releases. Their approach to the second problem is to search for an optimal tabular release. That is, a release that is decomposable and graphical which maximises some utility function whilst satisfying risk requirements. Specifically, they use a simulated annealing approach which is based on the fact that the space of decomposable graphical models can be traversed by the repeated addition / removal of single graph edges. This is a result due to Frydenberg and Lauritzen (1989).

The approach of Karr *et al.* is cautious, in that it assumes that a single user may have access to all the released tables. Although this is exactly the case for some statistical data, we are considering a situation where access to an online repository can be restricted to those with legitimate reasons for access. The approach of Karr *et al.* also requires that a suitable utility

function is chosen. But for data repositories that can only be accessed by registered users / user groups we propose a more flexible approach. If we assume that there will be no sharing of data between users, then we can allow users to access any data as long as the data satisfy risk requirements. In this case there is less incentive to impose a specific utility function on the user.

A General Disclosure Control Methodology

Our proposed general approach is based on the assumption that users who are carrying out scientific research, and who have been through appropriate vetting procedures, are less likely to share data than the public at large. So rather than treating the totality of the released data as a release to ‘the world’, we consider the risks posed by each group separately. Given that the data will generally be used for scientific research, we take the need for data quality seriously, and adopt the approach of releasing unperturbed data. Thus we have a similar situation to that of Karr *et al.* However, we feel it inadvisable to impose utility functions on the user. Each user will want to access specific data for specific purposes, and it seems more appropriate to allow users to decide for themselves which data are most useful. Only the group itself will be able to restrict access to high utility data by requesting low utility data.

User groups may be individual researchers, research groups, universities or wider groups such as ‘the media’. Some groups will be subject to stricter risk requirements than others. Specific requirements are based on the external information that is likely to be available to the users in a group and the type of attack on the data that such a user could reasonably mount. The main criterion for describing a group is that we should be very confident that a member of a group will not share data with a member of a different group.

The sets of tables to which a user can initially have access are defined by the risk requirement (and, of course, the data themselves). For example, the requirement might be that no zeros in the base table should be implied by the released tables. It follows from the bounds calculations for decomposable graphical releases (see Dobra and Fienberg, 2000) that any set of marginal tables that is graphical and decomposable and which contains no zeros is a releasable set. We also propose that the user should be afforded some limited opportunity to query the repository to investigate how future releases might be limited by a current query, before the query is actually submitted. This raises the question of how useful the information obtained from such an investigation would be to a determined data intruder, and how this activity might need to be limited. This question is part of ongoing research; as is the issue of analytical, rather than tabular, outputs.

It is not impossible that exact analytical outputs would allow an intruder to make reliable inferences about individuals. But rather than adopt the ‘perturb then analyse’ approach we feel that it is better to analyse the exact data, then limit or perturb the analytical outputs. For example, the precision of the output of an ANOVA table can be limited without altering any substantive conclusions that might be drawn from it. Analytical outputs might be releasable when the raw data they require could not be released.

Thus there are two possible types of output, tabular or analytical. Tabular outputs are easier to risk assess, but sometimes only analytical outputs will be safe. The general release architecture is shown in Figure 1.

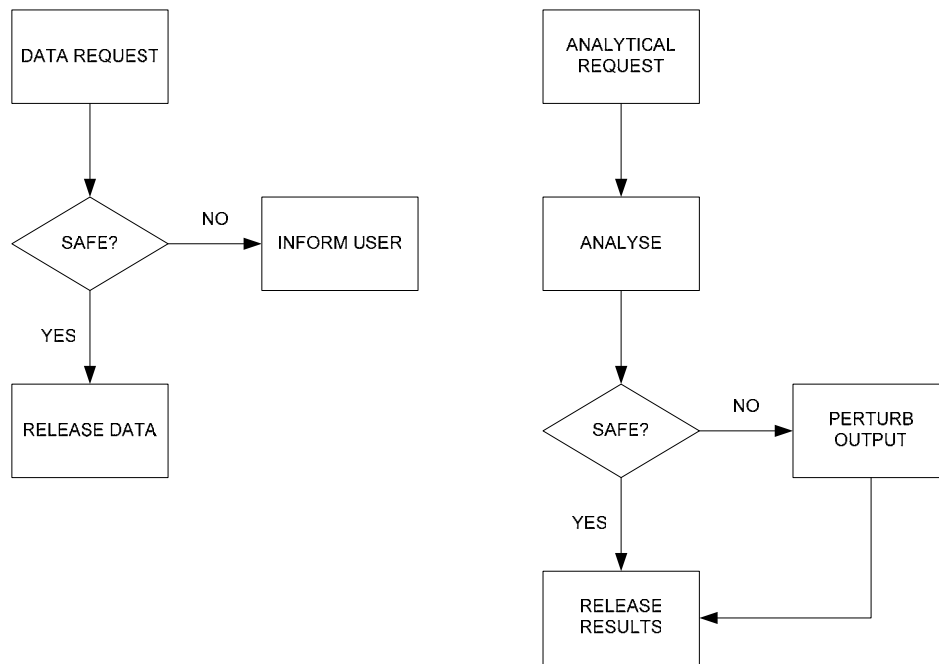


Figure 1. Proposed data release architecture.

Concluding remarks

This paper goes some way to outlining the issues and possible solutions for the disclosure risk issues associated with grid based medical data repositories. It also provides a demonstration of a very “mixed methods” approach to researching a complex data problem. Confidentiality and therefore disclosure control has legal, social policy, administrative, computational and statistical elements and the research outlined here touches on all of these. Maintaining focus on the problem at hand whilst incorporating these varied perspectives represents a challenge in itself. The extent to which we have met this challenge through an interdisciplinary team is one of the achievements of the research. The end result of this process is a functional model for disclosure control within medical data repositories and software (to be incorporated in CLEF) that instantiates that model.

The architecture and the general research method we have described here has general relevance for any remote data access system, and specific relevance to medical data.

References

- Caldicott Report (1997): London, Department of Health.
- Cayton, H. and Denegri, S. (2003): ‘Is what’s mine my own?’ *Journal of Health Service Research and Policy*, Vol. 8.
- Dobra, A. and Fienberg, S. E. (2000): ‘Bounds for Cell Entries in Contingency Tables given Marginal Totals and Decomposable Graphs’, *Proceedings of the National Academy of Sciences*, 97, No.22, pp.11885-11892.
- Doyle, P.; Lane, J.; Theeuwes, J. J. M.; and Zayatz, L. (eds) (2001): *Confidentiality, Disclosure and Data Access*, New York, Elsevier.

- Duncan, G. T. and Lambert, D. (1986): 'Disclosure-Limited Data Dissemination', *Journal of the American Statistical Association*, Vol.81, No.393, pp. 10-18.
- Elliot, M. J. (2005): 'Statistical Disclosure Control', *Encyclopaedia of Social Measurement*, Volume 3, New York, Elsevier. pp. 663-670.
- Elliot, M. J. and Dale, A. (1999): 'Scenarios of attack: the data intruders perspective on statistical disclosure risk', *Netherlands Official Statistics* 14.
- Frydenberg, M. and Lauritzen, S.L. (1989): 'Decomposition of Maximum Likelihood in Mixed Interaction Models', *Biometrika* 76, pp. 539-555.
- Kalra, D.; Singleton, P; Ingram, D.; Milan, J.; Mackay, J.; Detmer, D.; Rector, A. (2003): 'Security and confidentiality approach for the Clinical E-Science Framework (CLEF)', www.clinical-escience.org/industrial/sep2003/AHM2003-DKAlra-CLEF-Security-Confidentiality-Paper.pdf
- Karr, A. F.; Dobra, A.; Sanil, A. P.; Fienberg, S. E. (2002): 'Software Systems for Tabular Data Releases', *International Journal on Uncertainty Fuzziness and Knowledge-based Systems* 10(5), pp. 529-544.
- Kluge, E. H. (2001): 'Professional codes for electronic HC record protection: ethical, economic and structural issues', *International Journal of Medical Informatics*, Volume 60, Issue 2, pp. 85-96.
- Lauritzen, S.L. (1996): *Graphical Models*, Clarendon Press.
- Marsh, C. (1991): 'Privacy, confidentiality and anonymity in the 1991 Census' in A. Dale and C. Marsh (eds): *The 1991 Census User's Guide*, HMSO.
- NHS (2002): *Share with Care, Peoples Views on Consent and Confidentiality of Patient Information*, London, NHS Information Authority.
- ONS (2006): *Disclosure Review for Health Statistics, 1st Report for Guidance for Abortion Statistics*, London, ONS.
- Purdam, K. (2006): 'The Nation's Data', *Evidence and Social Policy* Vol. 2, No. 2.
- Purdam, K. and Elliot, M. (forthcoming, in press 2006): 'Data Quality and Utility', *Environment and Planning*.
- Rogers, J. (2005): 'Publicly Reported Breaches in EPR Confidentiality', CLEF Working Paper. www.cs.man.ac.uk/mig/people/jeremy/Confidentiality.html
- Skinner, C.J. (1992): 'On Identification Disclosure and Prediction Disclosure for Microdata', *Statistica Neerlandica* Vol. 46, No. 1, pp.21-32.
- Smith, D., and Elliot, M. (2003): 'An Investigation of the Disclosure Risk Associated with the Proposed Neighbourhood Statistics', *Report for the Office of National Statistics*.
- Smith, D. and Elliot, M. (2006): 'A Measure of Risk for Aggregate Data', (submitted).

Taweel, A.; Rector, A.; Kalra, D.; Rogers, J. (2003): 'CLEF Joining up Healthcare with Clinical and Post-genomic Research', www.health-informatics.org/hc2004/P31_Taweel.pdf

U.S Department of Commerce (1978): 'Report on Statistical Disclosure and Disclosure Avoidance Techniques', *Statistical Policy Working Paper 2*, Washington.