Statistical Disclosure Control and Protecting Privacy within a Clinical Data Warehouse

M. J. Elliot[1], D. Kalra[2], P. Singleton[2] and D. Smith[1]

Abstract

There is increasing interest in the development and research use of large scale repositories of clinical data, known as Clinical Data Warehouses (CDWs). CDWs facilitate the pooling of data from one or more electronic health record systems, providing a potentially useful resource for clinical research. Whether or not there is consent for the data to be used for research, there is almost always a need to protect the privacy of the individuals whose data is stored within the data warehouse. Formal identifiers such as name and address are usually not directly relevant to research questions and can be removed before data is released for research purposes. However, this alone does not preclude the identification of individuals within a CDW or the discovery of sensitive information about such individuals. By examining an example clinical data warehouse, developed during the Clinical E-Science Framework (CLEF) project, the authors have been able to consider these disclosure risks in detail. This paper describes the purposes of clinical data warehouses (CDWs), the risks posed through statistical disclosure, how these can be limited by appropriate use of statistical disclosure control (SDC) methods, and discusses how these can be applied in the context of clinical data warehousing.

Keywords: Statistical disclosure control, SDC, clinical data warehouse, CDW, privacy, confidentiality.

[1] University of Manchester
[2] University College London

1. Introduction

There is increasing interest in the development and research use of large scale repositories of clinical data. Sometimes these repositories arise as a direct result of the need to house the large scale datasets arising from prospective cohort or panel studies, and sometimes through the compilation of information from multiple sources of clinical data such as electronic health records. These repositories are often referred to as Clinical Data Warehouses. It is now widely recognised that important ethical and governance principles apply to such warehouses, as recently documented in the form of an ISO Technical Report (Grant et al 2006) and an ISO Technical Specification (Grant et al 2010). It is important that sensitive information about patients should not be inadvertently made available to third parties, a situation known as Statistical Disclosure. At an ICO workshop[3] in March 2011 on data anonymization, speakers from various backgrounds all acknowledged that 100% anonymization of any useful data was not a practical possibility, and therefore absolute guarantees of confidentiality are invalid. Instead, the appropriate framework in which to consider these issues is one of risk management. Although its potential within the field of health informatics is only now being explored; with official and social statistics, the theory and practice of Statistical Disclosure Control(SDC)[4] has been an active research area for many years (see Duncan et al (2011) for a recent overview).

By examining a sample clinical data warehouse, developed during the Clinical E-Science Framework (CLEF) project, funded by the MRC in two waves between 2002 and 2008, the authors have been able to consider the disclosure risks present in CDWs and apply statistical disclosure risk assessment methods to formally assess them. This paper summarizes the areas of risk identified (which also pertain to any similar clinical data warehouse) and proposes how statistical disclosure control methods might be used to assess and mitigate the risks.

1.1 What are CDWs and what are they for?

'Medical research' covers a wide range of activities, but is commonly associated with the analysis of experimental data resulting from clinical trials that assess the safety and efficacy of pharmaceutical drugs. Clinical data warehouses address the other end of the medical research spectrum: observational data, often routinely-collected, though sometimes collected as part of a long-term cohort study.

Clinical trials are often used to test a scientific hypothesis under carefully controlled conditions, which may be hard to reproduce in routine clinical practice, and are therefore

---

[3] http://www.ico.gov.uk/news/events/~/media/7B2FE4E551C84BBFA97E7A7F076219B3.ashx (Accessed 20th October 2012)

[4] SDC is also known as statistical disclosure limitation in some texts. The difference is transatlantic the two are synonymous.

rightly regarded as the gold standard for hypothesis testing. However, CDWs do have many valuable properties, which can provide a powerful complement to controlled trials:
CDWs are intended to provide large amounts of data on actual observed clinical phenomena, healthcare processes and their positive and negative outcomes.
The data contained in CDWs allow the testing of hypotheses without the creation of an experiment (with all the associated direct costs).
Ethical considerations might mean that conducting clinical trials can be problematic, and so some hypotheses can only be investigated via observational data.
Observational data reflects actual practice rather than the laboratory conditions of a randomised clinical trial (RCT) so may be more pertinent for answering questions about best practice rather than scientific fact.
Although analyses of routine observational data are subject to confounding and various sources of bias to a greater degree than clinical trial data, sample sizes can be much larger than is feasible in clinical trials giving greater statistical power.

Epidemiologists and other analysts use data-mining techniques on CDWs to extract links and relationships between aspects of the health records, perhaps to discover relationships in the data which may not have been suggested through formal deductive processes. Examples of such discovery processes include the identification of the predictors of recovery following: brain injury, hip fracture, cardiovascular risk in haemodialysis patients and the monitoring of asthma.
The data contained in CDWs can be used to conduct historic cohort studies or case-control studies. A significant disadvantage with prospective cohort studies is the amount of time required to collect the data, potentially recording information about exposures, outcomes and other relevant variables over several decades. A CDW might already contain suitable data, enabling a historic cohort study to be carried out whilst avoiding the time and expense of data collection. Alternatively, a CDW could be used to identify disease cases and controls for a case-control study. This might enable population, rather than hospital-based, studies to be carried out. A large database might allow better matching of controls to cases, and a larger number of disease cases in the case of rare diseases. Observational studies facilitated by CDWs are less scientifically rigorous than well-conducted randomized controlled trials (RCTs) and the former should not be regarded as a substitute for the latter. They are subject to various forms of bias and the possibility of confounding. Yet, ethical considerations might dictate that only observational studies can be performed. A CDW offers the potential for observational studies to be carried out quickly and cheaply, without any obvious disadvantages that are not already characteristic of these types of study. In some cases observational analyses of data contained in CDWs might be a useful precursor to developing an RCT; a medium for exploratory analysis helping to identify the problem area and to generate potential hypotheses and appropriate methods for testing those hypotheses.

Examples of general-purpose CDWs are:

General Practice Research Database (GPRD)
The GPRD is the world's largest computerized database of anonymized longitudinal medical records from primary care that is linked with other healthcare data. Currently, data are

being collected on over 3.6 million active patients (approx. 13 million total) from around 488 primary care practices throughout the UK. It is the largest and most comprehensive source of data of its kind and is used worldwide for research by the pharmaceutical industry, clinical research organisations, regulators, government departments and leading academic institutions.[5]

Q-Research

QRESEARCH is a large consolidated database derived from the anonymized health records of over 12 million patients. The data currently come from 602 general practices using the EMIS clinical computer system. The practices are spread throughout the UK and include data from patients who are currently registered with the practices as well as patients who may have died or left. Historical records extend back to the early 1990's making it one of the largest and richest general practice databases in the world.[6]

UK Biobank

UK Biobank aims to study how the health of 500,000 people, currently aged 40-69, from all around the UK is affected by their lifestyle, environment and genes. The purpose of this major project is to improve the prevention, diagnosis and treatment of a wide range of illnesses (such as cancer, heart disease, diabetes, dementia, and joint problems) and to promote health throughout society.[7]

IMS Health

IMS LifeLink™ describes itself as a unique global program of patient-centred information, analytics and consulting. Through LifeLink's longitudinal disease and treatment dynamics, healthcare stakeholders gain critical knowledge about drug utilization, prescribing and cost-of-care trends. LifeLink offerings range from syndicated products and services infused with new patient-level metrics to proprietary custom analytics and consulting engagements, all incorporating a patient perspective. Building on significant IMS investments in technology and anonymized patient-level data around the world, the underlying LifeLink database is sourced from longitudinal prescriptions, health insurance claims and electronic medical records.[8]

Dept. of Health Hospital Episode Statistics database

HES is the national statistical data warehouse for England of the care provided by NHS hospitals and for NHS hospital patients treated elsewhere. HES is the data source for a wide range of healthcare analysis for the NHS, government and many other organisations and individuals. HES came about in 1987 following a report on the collection and use of hospital activity information published by a steering group chaired by Dame Edith Körner. For example, the total number of hip replacements can be found by searching the database for records that contain the appropriate procedure or intervention codes. In this case the

---

5 www.gprd.com/home/default.asp

6 www.qresearch.org/Public/WhatIs.aspx

7 www.ukbiobank.ac.uk/docs/Informationleaflet130608.pdf

8 www.pharmavoicemarketplace.com/featured.php?company=75058&m=f

resulting figure (many thousands per year) reveals nothing about the individuals concerned, and may be freely publicized.[9] Now transferred into CfH SUS – see below.

NHS CfH Secondary Uses Service

Secondary Uses Service (SUS) is the single source of comprehensive data to enable a range of reporting and analysis. The data currently managed within SUS is derived from the commissioning datasets, which providers of NHS care must submit and make available to commissioners. In future, wherever possible, data will be captured automatically from NHS operational systems including the NHS Care Records Service and other National Programme for IT services including Choose and Book, the Patient Demographics Service and the Electronic Prescribing System. SUS provides a range of software services and functionality which enable users to analyse, report and present this data. It is the single, authoritative and comprehensive source of high quality data. It provides a secure environment that maintains patient confidentiality to national standards.[10]

CDWs may be more constrained by condition, geography, or research focus:

Avon Longitudinal Study of Parents and Children (ALSPAC)

ALSPAC is a long-term health research project. ALSPAC recruited more than 14,000 pregnant women with estimated dates of delivery between April 1991 and December 1992. These women, the children arising from the index pregnancy and the women's partners have been followed up since then and detailed data collected throughout childhood. ALSPAC is a two-generational resource available to study the genetic and environmental determinants of development and health.[11]

Whitehall study

The Whitehall study examined mortality rates over ten years among male British Civil Servants aged 20-64. The study was an attempt to avoid some of the problems created by the use of general social class groupings, for example, the heterogeneity of occupations within a single class leaves room for multiple interpretations. The Whitehall study concentrates on one "industry" in which there is little heterogeneity within occupational grades and clear social divisions between grades. A second longitudinal study of British Civil Servants (Whitehall II) was initiated to investigate occupational and other social influences on health and disease. The final sample was 6900 men and 3414 women aged 35-55 in the London offices of 20 civil service departments.[12]

Clinical trials may generate quite large medical databases, but these are usually constrained to a single research topic. Although this does not preclude their re-use in other research projects, these are usually in very closely related areas. We take it that data within a CDW is intended and designed to answer a wide range of possible enquiries.

---

9 From www.hesonline.nhs.uk

10 www.connectingforhealth.nhs.uk/systemsandservices/sus/background

11 www.bristol.ac.uk/alspac/sci-com

12 www.workhealth.org/projects/pwhitew.html

1.2 The CLEF project

The CLEF project (Clinical E-Science Framework, MRC GR/M54919, 2002-2005) was a major UK e-Science project which developed a federated clinical data warehouse architecture to integrate data from diverse health record systems and support the design and execution of complex research queries across populations of patient records. The design and implementation of data protection, de-identification and technical security policies was an important aspect of the work. CLEF-Services[13] was a follow-on successor project to CLEF. The goal of this follow-on project was to investigate the barriers to, and challenges for, the secondary use of operational clinical data in bio-science research, with a particular focus on disclosure control.
CLEF established a clinical data warehouse by taking cumulative extracts of data from the electronic health records held by the Royal Marsden Hospital. Ethical approval was granted for the use of the records of almost 22,500 deceased patients, within strict confidentiality and data security protocols which included permission for research specifically on the confidentiality approach. The CLEF CDW was fully instantiated and populated, and used by the approved project partners at University College London, the University of Manchester, the University of Sheffield and the Open University. It was never made available to outside teams to be used as an operational research tool. (The project was set up to develop methods and best practice, not to deliver an operational service for research. The CDW has therefore now been decommissioned.) The fully populated repository contained over 12 million nodes (4 million headings and subheadings, and 8 million actual patient observations ranging in complexity from single numeric values to whole word processed letters and test reports).

The range of clinical departments and specialities contributing data to this CDW are listed below.

- Basic patient registration details, episodes and length of stay
- Clinical diagnosis, specific details of the cancer being treated
- Clinical chemistry test results
- Haematology test results
- Cytology investigation results
- Radiology investigation results
- Histopathology test results
- Theatres: operation notes and procedure details
- Pharmacy (medications and chemotherapy administration)
- Radiotherapy treatment details and reports
- Case notes: outpatient clinic letters and discharge summaries
- Death Certificate information

[13] MRC G0300607, 2005-2008

The focus on cancer diagnosis, assessment and treatment records was a deliberate choice to limit the complexity of creating this CDW and analysing its content, while still providing a meaningful and valuable subject area. The data supplied by the Royal Marsden was de-identified at source by removing the conventional demographic information, scrambling all formal identifiers and also removing occurrences of the each patient's names from within free text fields such as clinical letters and reports.

2. Statistical Disclosure Control

Disclosure risk is a complex topic with its own research field, conferences and journals. We do not attempt here to capture all of the possibly relevant features but to illuminate some of the key features which informed our decision making. The interested reader is referred to Willenborg and de Waal (2001) or Duncan et al (2011) for overviews.

SDC revolves around a set of conceptual dichotomies; the key ones for the current purposes are:

Identification v. attribution
Microdata v. aggregate data
Utility v. risk
Sample data v. population data
Logical inferences v. probabilistic inferences[14]

2.1 Identification and Attribution

Statistical disclosure has been defined as the "accurate attribution of information about a population unit to that population unit", Elliot (2005). There are two key processes in statistical disclosure:

Identification - the association of a population unit (e.g. a person) with a particular data unit[15]

Attribution - the association of information in a dataset with a particular population unit.

Identification generally follows from a *data intruder* being able to link known information about a particular population unit usually referred to as a *target* with information in a database (often simply referred to as a file). This is achieved through matching on a set of *key variables*. An *identification file* containing *formal identifiers* (e.g. name, age and address) could be used to identify an individual population unit. Matching the identification file

[14] The last pair are sometimes referred to as "exact" and "approximate" inferences. Thus "exact attribution" would be the discovery of new information regarding a target that the intruder could know with certainty.

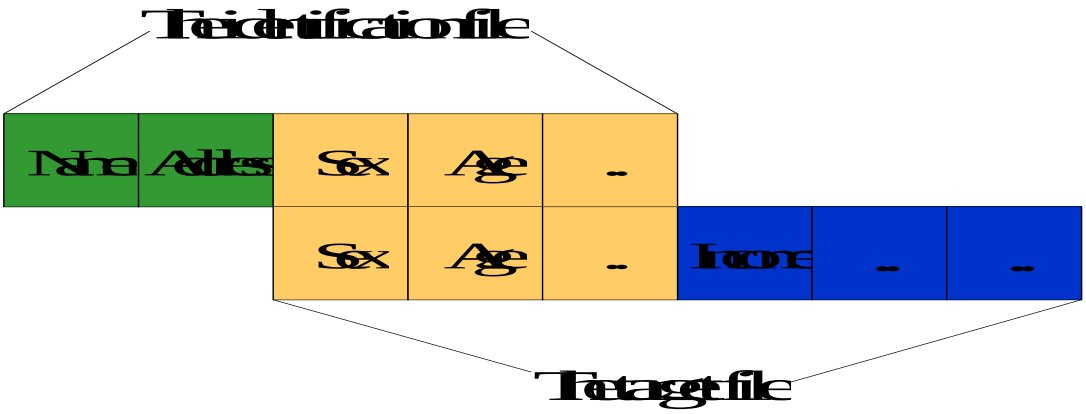
[15] A Data unit is a record in a dataset which corresponds to a population unit (person, household business etc). So data units are data representations of population units which are actual entities.

against a *target file* (a database containing no formal identifiers that contain both the key variables and information on *sensitive* variables) would allow sensitive information to be associated with the target. This process is illustrated in Figure 1.

The known information might stem from personal knowledge of the target, in which case no identification file is needed. In this case the key variables are those that might be known to the intruder, rather than those which are common to an identification file and a target file.

Anonymization (by which we mean here simply the process of removing formal identifiers[16]) makes it more difficult to associate data records with population units. However, it might still be possible if a file contains sufficient detail on the key variables. A very specific value for "occupation" such as "Mayor of London" would allow the relevant individual to be identified with absolute certainty. This illustrates why limiting the level of detail is an important aspect of disclosure control.

Much of the SDC literature on identification risk concentrates on *uniques*: records within a file that have no matching records within that file. Records that are unique on small subsets of variables are generally of greater concern, particularly if they are unique on *key variables* (variables which are common to two datasets and therefore could be used to link those datasets – e.g. the yellow boxes in figure 1) [17]. Without unique records it is impossible for an intruder to associate a population unit with a record with certainty. However, this does not prevent attribute disclosure.



[16] Anonymisation is - unfortunately - used variously in the literature and in policy document. Elliot and O'Hara (in prep) refer to the technical definition that we use here as Anonymisation-1 and it that definition that we will employ consistently throughout this paper. Elsewhere it used to mean data to which statistical disclosure control has been applied (Anonymisation-2). In other cases it means it is not possible to identify people in this dataset (anonymisation-3).

[17] Note that, in this context, you can only define variables as keys by reference to two (or more) datasets. This is different from the standard use of the term in database parlance where it is taken to mean a unique identifier variable often a serial number.

Figure 1. An illustration of the key variable matching process leading to disclosure from Elliot (2001).

Attribute disclosure occurs when an intruder associates information with a target. Again, it depends on matching known information against information within a published file. Figure 1 demonstrates how the values for sensitive variables such as "income" can be associated with an individual identified via matching on a unique record. However, attribution does not require a unique match. If the set of matching records contains a variable with a common value, then that value can be associated with the target. Attribution could be viewed as a process of information matching, rather than record matching.

It should be clear from the above discussion that identification does not imply attribution; and attribution does not imply identification. A record might only be unique with respect to the set of all the variables within a target file. Thus identification would require knowledge of the complete record, and no new information would be disclosed by associating a target with the record. Identifying several possible matches within a file which all had low income would allow the inference that the target was on a low income.

It can be argued that it is attribution that causes disclosure, rather than identification. In a very real sense this is true. Yet the possibility of a believable claim of disclosure is often also of concern to Data Stewardship Organisations (DSOs)[18]. Identification is sometimes considered to constitute disclosure, even if it does not lead to attribution. The SDC literature is rather heavily weighted to identification risk.

2.2 Microdata and Aggregate Data

In SDC there is usually a distinction made between microdata and aggregate data. Microdata are in the form of a collection of individual records, whereas aggregate data are often in the form of a table containing a count for each possible combination of variable values. [19]
A microdata file might contain personally identifiable information such as names and addresses. It might also contain variables on a continuous scale with variable values

[18] Following Duncan et al (2011) we define DSO's to be organisations that have the twin objectives of protecting entrusted data by providing confidentiality and assuring its beneficial use by researchers and policy analysts.

[19] Another form of data is magnitude data, although it is far less relevant to CDWs than microdata and aggregate data and will not be discussed. Magnitude data is in the form of totals or average values, e.g. turnover. The disclosure risk is that a published industry total would enable a large firm to put upper bounds on the turnover of competing firms by simply subtracting its own turnover from the industry total. The (n, k) rule (Willenborg and de Waal, 2001) specifies that if any n units contribute more than a proportion k of the total, then the data are insecure. There is also a p% rule, where data are considered to be insecure if they would allow any respondent values to be estimated within p% of their true value.

specified to high levels of precision. Thus, in a microdata file it would not be unusual to find many unique records, particularly if records contained a large number of variables as is usually the case with most healthcare data.

In contrast, data that are provided in aggregate form generally do not include variables on a continuous scale or that have large numbers of possible values (such as names and addresses).

Specific SDC methods will usually be predominantly associated with either microdata or aggregate data. A microdata file and an aggregate table could contain exactly the same information.

2.3 Utility and Risk

Data are generally made available because they are considered to be useful for some purpose. The potential benefits of CDWs were briefly discussed earlier. Some data are considered to be sensitive. Many people would rather that the general population did not know their earnings. Although sensitivity is subjective, and will differ across geographical and cultural boundaries, it is important to recognize that the goal of SDC is to allow useful data to be released whilst protecting sensitive data. Thus there is a trade-off is between utility and risk, where risk has to be assessed both in terms of the probability that information will be disclosed to a data intruder, and the sensitivity of the disclosed information. Just as variables tend to be dichotomized into key and non-key variables, they also tend to be dichotomized into sensitive and non-sensitive variables. This could be viewed as an over-simplification. After all, it is impossible to know exactly what information a data intruder will have regarding a target, and it is often specific levels of a variable that are sensitive rather than all the levels of a variable. Such simplifications are generally necessary to produce workable SDC methods.

In some cases sensitivity might stem from inferences that could be made on the basis of a disclosed variable level rather than the disclosed level itself. A negative HIV status might be considered to be less sensitive than a positive status, although the existence of either status would imply a test and possible membership of a high risk group.

2.4 Risk assessment

The final two dichotomous concept pairs are sample data v. population data and Logical inferences v. probabilistic inferences, both of which are critical to risk assessment.

The distinction between sample and population data is very important. The above discussion relating to record matching and disclosure predominantly applies to population data (the exceptions being structurally uniques (for example being the Mayor of London as

ones occupation which is, which identifies the individual in either sample or population data). Sample data generally disguise the true number of matches to a targeted individual within a population. Thus even a unique match to a data unit a sample data-set could lead to an incorrect identification if it were not unique in the population data-set. There is still an issue with structural uniques and potentially with sample uniques that could be considered to be so unusual that they are very likely to be population unique (the canonical example of this is the 16 year old widow). However, in many circumstances sample data can offer a substantial degree of protection against disclosure whilst not unduly impacting upon data utility. For example, 90% sampling has been used to good effect by Sparks et al. (2008) in the controlled remote access setting.

The previous examples have related to logical inferences. In practice there is generally some uncertainty over the correctness of published data, so even inferences based on logic might be wrong. For the purposes of risk assessment this potential source of uncertainty is generally ignored. In addition to logic, a data intruder might make probabilistic inferences. Methods such probabilistic record linkage (Fellegi and Sunter, 1969) might yield probabilities of correct matches that are high enough to be of concern.

Risk assessment often starts with an "attack scenario" (Elliot and Dale, 1999). An attack scenario describes both the type of intruder and the strategy employed by the intruder to recover information. For example, an intruder could be a member of the public who knows that a data record relating to a neighbour is contained in a published database. The intruder might attempt to identify the relevant record by matching against known information about the neighbour. Once an attack scenario has been identified the risk can be assessed by adopting the position of the data intruder and attacking the data (Paass 1988) (Mokken et al., 1992). The output of such an attack is often a risk measure such as the SAP measure for attribution risk with aggregate population data (Smith and Elliot, 2008). Others have developed rules which are designed to classify datasets according to risk.

The relevancy of a risk measure / criterion is dictated by the form of the data and type of disclosure that is of concern. These are generally contained within the description of an attack scenario.

2.5 Disclosure Control Methods

If a dataset is found to be too risky according to some suitable criterion, then disclosure control methods might be applied. These methods are designed to reduce the risk of disclosure whilst allowing useful data to be released.

Suppression is an important form of disclosure control. Duncan et al. (2011) describe suppression as the "denial of data instances". They discuss the deleting of records and variables in microdata, which they respectively term *record suppression* and *attribute suppression*. Record suppression might be undertaken to remove unusual individuals who

might be easily identified. Attribute suppression might be used to remove formal identifiers, key variables or sensitive variables. Thus sampling and anonymization are actually forms of suppression. Each time a DSO decides to release aggregate data on a subset of the variables for which they have data, it is *de facto* a suppression of the unreleased variables.

0	3	0	1	4
0	1	0	0	1
4	1	0	2	7
1	0	7	1	9
5	5	7	4	21

Figure 2. *Data release with a risky cell*

Figure 2 shows a possible aggregate data release for two variables. The count of 1 in the second row and second column might be considered to be too risky. For instance, the second level of the row variable might be unusual, enabling identification of the relevant individual for whom the second level of the column variable could then be inferred.

0	3	0	1	4
0	*	0	0	*
4	1	0	2	7
1	0	7	1	9
5	*	7	4	*

Figure 3. A suppressed cell with suppressed marginal counts

Figure 3 shows how the risk of disclosure can be reduced by suppressing cell values. (This is simply record suppression, but with the data in aggregate form.) The effect of suppressing the unique is to also suppress certain marginal totals. The nature of the data (counts) implies a lower bound of zero on the suppressed cells. There are no finite upper bounds. In the interests of data quality an attempt might be made to choose cell suppressions that allow the correct marginal distributions to be published.

0	*	0	*	4
0	*	0	*	1
4	1	0	2	7
1	0	7	1	9
5	5	7	4	21

Figure 4. A suppressed cell with preserved marginal counts

Figure 4 shows one possibility, of several, for suppressing cells so that the risky cell is protected and the publication of marginal counts does not allow the value of the suppressed cells to be recovered. In this case the risky cell can only be a 0 or a 1, so there is a lower degree of protection. In fact, if the other suppressed cell in the second row happened to be a structural zero, then the suppressions in Figure 4 would offer no protection at all. The original table could be inferred from the published counts. Finding a set of cell suppressions that adequately protects the data whilst maximizing data utility is known as the complementary cell suppression problem (Fischetti and Salazar, 1999; Salazar, 2008).

0	4	0	1	5
4	1	0	2	7
1	0	7	1	9
5	5	7	4	21

Figure 5. *A recoded data release*

Aggregating the first two levels of the row variable is an alternative approach to protecting the data in Figure 2. Aggregating variable levels is known as *recoding*. This suppression of detail can allow exact counts to be published.

At this point it is worth noting that attribution depends on the presence of zeros. . There are seven individuals who have the column variable at the third level. They all share the same value for the row variable. Identifying any of these individuals in the population would result in the disclosure of the applicable level of the row variable (assuming it was not already known to the intruder for the identified individual). None of the above measures taken to protect the marginal unique for the row variable protect these seven individuals. Identifying any one of a group of seven will generally be easier than identifying a single individual,

although the fact that all members of the group share a common level for the row variable might suggest that the inference corresponds largely with prior expectations. Thus the inference might not be particularly noteworthy and might be less likely to result in a claim of disclosure. This serves to demonstrate how difficult it can be to balance risk and data utility. Even identifying the risky cells can be a non-trivial exercise.

Suppression is the act of not releasing information, and that information can be quite arbitrary. It is not limited to records and variables. Statistical outputs might be suppressed in order to prevent certain inferences regarding the analysed data. The total earnings within an industry might be suppressed if there were only 2 companies involved, to prevent each discovering the earnings of its rival by simply subtracting its own earnings from the total.

Suppression contrasts with perturbation; the act of changing data before it is released. Data swapping is the process of swapping part of a record with another record. Any attempts to match records are faced with the issue that the levels of variables used for matching might be incorrect, and any inferred levels of other variables might be incorrect even if the match is correct. Data swapping preserves the univariate marginal distributions. Schemes might be implemented so as to preserve multivariate marginal distributions by requiring sets of variable values to be swapped jointly (between the same pairs of records). An analogous (essentially identical) method for aggregate data is controlled tabular adjustment (Cox et al., 2004). Cell counts are changed to protect risky cells, and complementary changes are made to non-risky cells in order to preserve certain marginal distributions.

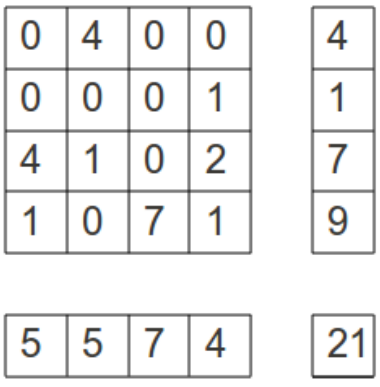


Figure 6. *Controlled tabular adjustment*

Figure 6 shows a possible solution to the table in Figure 2. The pattern of changes follows the pattern of suppressions in Figure 4. In this case the counts are changed, rather than suppressed, in a way that maintains the marginal totals. Identification based on the row margin of 1 would lead to an incorrect inference regarding the corresponding value of the column variable. The purpose of disclosure control is not generally to guarantee that such deductive inferences are incorrect; it is to add sufficient uncertainty to such inferences

whilst not unduly impacting the statistical qualities of the data. Thus it is important that the data intruder knows that published data have been disclosure limited. This can also be valuable information for the legitimate analyst who might be able to adjust for the increased uncertainty.

A very basic form of perturbation for aggregate data is Barnardisation, adding or subtracting 1 from selected non-zero counts in aggregate data. Non-zero counts are selected for perturbation with probability (1-*p*) and are subsequently rounded down or up with equal probability. The scheme is described by the following conditional probability mass function,

$$P(j|i) = \begin{cases} p & if j = i \\ \frac{(1-p)}{2} & if j = i-1 \\ \frac{(1-p)}{2} & if j = i+1 \\ 0 & otherwise \end{cases}$$

where $P(j|i)$ is the mass function for the published count *j* given the true count *i*. and *p* is the Barnardisation parameter; the probability that a non-zero true count will be published (unperturbed).

The scheme is unbiased in the sense that the expected value of each Barnardised count is equal to the original count. It is not unbiased in other senses. For instance, original zeroes will not be perturbed whilst some ones might be rounded to zero. Thus Barnardised data sets tend to zero-inflated.

Other types of rounding scheme can be similarly characterized by probability mass functions (with probabilities of 0 and 1 for deterministic schemes). However, rounding cells independently does not preserve marginal counts and such schemes apply general protection rather than focussing on risky cells. They are becoming less frequently used. Where they have been used the schemes tend to be functions of one or two parameters. The parameters are generally suppressed to increase the level of protection. This implies an additional source of risk; the accidental release or discovery of the parameters.

Although suppression and perturbation seem to be distinct concepts, there are SDC methods which do not fall cleanly within the definition of either. For instance, sensitive data might be suppressed and then replaced by values generated using multiple imputation methods (see for example Reiter, 2005). The resulting “synthetic” data are then released. Although the original values are suppressed, this is a form of suppression that takes place every time a data value is perturbed (suppression of the original data value). However, the

published value does not depend on the original value at all. It depends on the other values in the data and the multiple imputation method used. Thus it is not perturbation in quite the same sense as, say, Barnardisation.

2.5 Pre-tabular and post-tabular methods

Data are often released as (aggregate) tables. These are generated from an underlying database. The process generally involves anonymization and the categorization of continuous variables. It might also involve the suppression or recoding of certain variables for disclosure control purposes. The DSO can apply perturbative SDC methods such as data swapping to the underlying database before the tables are generated. This is known as pre-tabular disclosure control. Alternatively a DSO might choose to apply methods such as Barnardisation to the generated tables. This is known as post-tabular disclosure control. Of course, a DSO could choose to use both. For the 2001 UK census data swapping was applied to the raw microdata before tables of counts were generated for small geographical areas. A post-tabular rounding scheme was then applied to low counts in those tables.

Pre-tabular methods ensure that released tables are consistent (assuming no additional post-tabular disclosure control). Distinct tables with common variables will have equal marginal distributions for those variables. This can make data analysis easier. However, the level of protection required for a whole database is generally much more than the level of protection needed for tabular releases involving a small number of variables (which additionally benefit from attribute suppression). So, small tabular releases will be subject to large degrees of uncertainty.

Two commonly discussed pre-tabular techniques are *k*-anonymity and *l*-diversity. Samarati and Sweeney (1998) introduced *k*-anonymity. A dataset satisfies *k*-anonymity, for $k>1$, if at least *k* records exist in the dataset for each observed combination of key variable levels. Thus no unique match can be made against any particular data unit on the key variables. It does not generally protect against exact attribution because data units in the same k-group might have the same values on sensitive (Smith and Elliot, 2008) (Domingo-Ferrer and Torra, 2008).

Machanavajjhala et al. (2006) introduce *l*-diversity. In its simplest form *l*-diversity requires that there must be at least *l* distinct values for each sensitive variable, for each combination of key variable levels. Thus it protects against exact positive attribution, but not negative attribution (the disassociation of a variable level with a target). Other forms of *l*-diversity are discussed in Domingo-Ferrer and Torra (2008).

Post-tabular methods can focus on the risks associated with the tabular release in question, but they can be inconsistent. This also provides an intruder with a means of attacking the data. Rounding schemes such as Barnardisation imply a set of linear constraints on table counts. In some cases these constraints can be used to recover the original counts. Take the

release in Figure 2 and the unique in the second row and second column. Barnardisation might leave this unchanged, implying that the original count must have been 1 or 2. If the corresponding row sum had been rounded down to 0, then the row sum would have to have been 0 or 1. This would allow the recovery of all the original values in the second row, allowing exact attribution to take place.

2.6 Safe settings and servers

The preceding discussion has related to the release of data in either microdata or tabular form. In many cases the DSO decides which data are safe to release and makes them generally available, often via a server that can be accessed via a web site. The UK data archive provides access to microdata from several social and economic data sets[20]. American FactFinder[21] provides access to aggregate data for population, housing, economic, and geographic variables in the United States. There are other options for a DSO.

In some cases it is possible to give access to data in a "safe setting". Approved analysts will be allowed to perform analyses without being provided with the raw data. This would sometimes require the user to travel to a particular location where the data could be analysed on a dedicated computer. An alternative is to allow users to access a server remotely to perform analyses. The computer is then referred to as an "analysis server". This is certainly more convenient, but implies additional risks associated with having the data on a networked computer. An alternative is to allow aggregate data to be distributed to approved users via a *table server*.

Analysis and table servers consider requests which might be refused. Thus there is a distinction between a table server and the release of predetermined aggregate outputs via a web site. The risk and utility of a requested output can be considered in the light of previously released outputs. The DSO can also take into account the trustworthiness of the requester. Risk assessment can be manual or automatic. Automatic systems have the obvious advantages of being able to serve requests in a more timely fashion.

There has been a move towards the server options as the necessary enabling technology has developed. These were the options considered for the CLEF repository.

3. Disclosure Control for the CLEF data

An analysis was made of the kinds of clinical research queries that might be made on a warehouse such as the CLEF repository. From this, it was possible to ascertain the kinds of result sets that might be requested (as a set of patterns, and example-specific instances of result sets). The disclosure risk research also included consultation with users regarding

---

[20] www.data-archive.ac.uk - accessed 12/9/2012

[21] factfinder2.census.gov – accessed 12/9/2012

analytical requests, the development of attack scenarios and the statistical assessment of disclosure risk.

Statistical Disclosure Control methods that had originally been developed for controlling disclosure risk in the UK Census and survey data were applied to these patterns of clinical research result set, and an SDC software tool was developed to perform risk assessments automatically. It was particularly noted that the potentially poorer data quality of routine clinical data actually protected against disclosive inferences.

The data release approach developed under the CLEF project is described in detail in Elliot et al(2008). The philosophy behind the approach was to draw on the positive aspects of many of the orthodox methods of data access and disclosure control outlined above whilst providing the user with the flexibility needed to meet their own analytical requirements. A combined table server / analysis server architecture we employed (see Figure 7). The key idea is that users can request unperturbed aggregate data, and if this is considered unsafe, then an analytical release can be requested. Analytical releases (statistical outputs) might be perturbed to guard against disclosure, but any such perturbations are made without significantly altering the substantive conclusions that could be reasonably made.

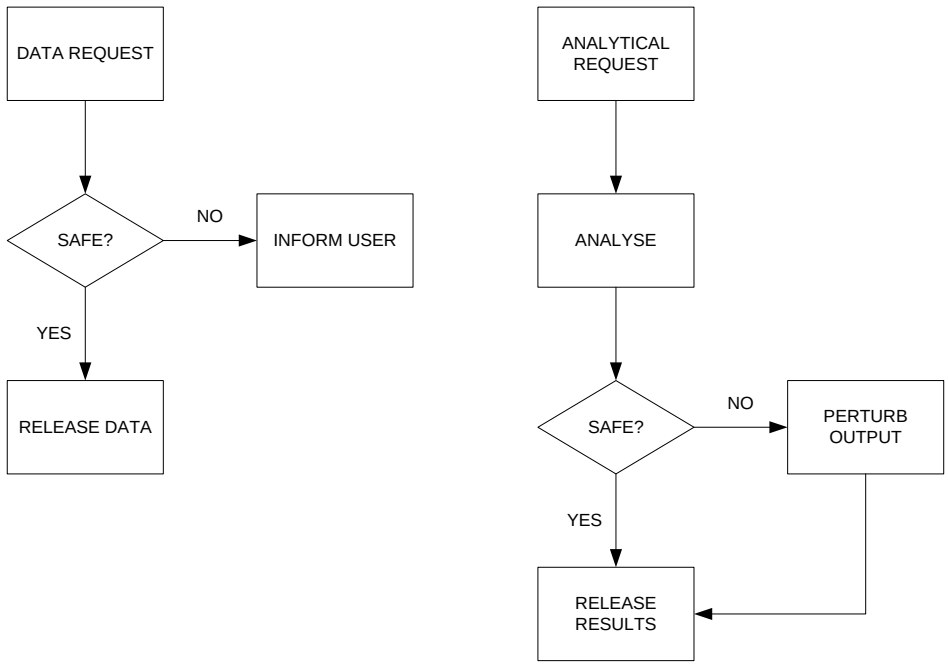


Figure 7. CLEF SDC architecture

The table server component was fully implemented in the project as it differs in some respects from existing servers. The analysis server component was not fully implemented as it does not fundamentally differ from existing servers.

'Users' of the system might be individual researchers, research organisations or less well-defined groups such as 'the public' or the 'media'. Different users are associated with different levels of trust. Trust relates to both to the likelihood that the user will attempt disclosive inferences and to the likelihood that the user will share the data with others (who might attempt disclosive inferences). For instance, a research organization might be provided data on the assumption / assurance that data will only be shared with other members of the same organization. Any data released to members of the general public is assumed to be shared by all. Thus disclosure risk is assessed on the basis of all the data released to a given user and all other released data that might be shared with the user. A researcher would generally have access to data that would not be released to the general public. Risk is assessed in terms of the cell bounds that can be placed on cells in the most detailed table (cross-classification over all available variables). As demonstrated earlier, the release of marginal tables places linear constraints on these cell bounds.

Karr et al. [2] considered the computational issues of assessing risk for marginal tables (cross-classifications on subsets of variables) through the generation of cell bounds. In general these bounds are not efficiently computed, making risk assessment difficult. However, there are efficient algorithms for computing bounds for data releases that have a certain type of structure [3]. The approach of Karr et al. is to identify a safe data release (set of marginal tables) that offers maximal data utility. They restrict consideration to releases for which the bounds can be efficiently calculated. Requests for tables which are not a subset of this “full release” are refused. The benefit of this approach is that it avoids situations where low utility releases subsequently limit the release of higher utility releases.

In practice there might be many sets of tables that satisfy a chosen risk threshold. Some might be subsets of larger releases that also satisfy the threshold. These are not candidates for a full release as they necessarily do not contain data that could be released safely in addition to the tables they contain. So the candidates for the full release are the maximal sets of tables that satisfy the threshold, where maximality is in the sense of having no superset of tables that also satisfies the threshold. A trivial example would be the bivariate case where both the 2-way table and the release consisting of the two 1-way margins did not satisfy a risk threshold, yet the two releases consisting of the 1-way margins each satisfied the threshold. The full release under the approach of Karr et al. would be the 1-way margin that maximized data utility.

The main difference with the CLEF table server is that not all releases are considered to be to the world. Approved individuals or organisations are trusted not to share data, and so each can choose to request the data that they consider to be most useful. Thus no utility function is imposed upon them. In the trivial example above the user would be able to request either 1-way margin, but would not subsequently have access to any more data. Over time the data released to a user might approach a maximal safe release (given the risk measure and threshold relevant to that user) and further data might not be provided. But in the meantime the user has had the opportunity to, in effect, select the maximal release that maximizes their data utility.

This still leaves the issue that releases to less trusted users might limit what can be released to more trusted users. Again the trivial example illustrates the problem. The release of a 1-way margin to, say, the media (for possibly trivial purposes) would prevent the other 1-way margin being released to a research organization for legitimate research purposes. A pragmatic way to deal with is to be extremely conservative regarding releases to less trusted users until a sufficient number of more trusted users, such as researchers, have queried the system. In this way it tends to be the high utility research releases that limit the lower utility public or media releases. However, the CLEF table server also allows searches for maximal safe releases for a given risk criterion based on cell bounds in the detail table. Although searches are not guaranteed to find all such releases over a dataset with many variables, the search does help to identify individual tables that are contained in many maximal safe releases. These are less likely to limit the options for researchers if released to less trusted users, and can be released to less trusted users (as long as they are safe by the risk criteria relevant to those less trusted users). Thus the system is able to intelligently manage the release of data to different classes of user whilst respecting risk thresholds and attempting to maximize overall data utility.

4. Discussion

It is important to balance the risks against the operational costs and possible adverse impact on research outcomes. If we set confidentiality requirements too high, we are likely to de-bar certain very worthwhile research projects; if we are too slack then patients will quite rightly ask for their data to be excluded at source – eliminating certain forms of research altogether. If we impose high confidentiality and security costs, then some research will simply cease to be cost-effective (at least as research, even though the benefit from the research might be huge) – see Singleton & Wadsworth (2006) on costs of gaining consent by different methods. The possible benefits from CDWs will simply vanish if the set-up costs become too great.
There is no doubt that there needs to be some requirement to limit access on queries; unfettered access to the data means that the information could be mined inappropriately. These checks can be either ante hoc or post hoc, though the former is clearly to be preferred. If the latter, then checks need to be thoroughly policed and any transgressions actively pursued. Sweeping matters under the carpet is clearly not acceptable, nor is 'See no evil…'.

It is possible to attempt to gauge risk on a query-by-query basis, though it is difficult to generalize as the risks tend to be specific to the data set and the level of access granted.

Ideally, most access can be handled within the general structure for the vetting of queries described above. However, there may remain a need to review access to high-risk data items on an *ad hoc* basis. This is particularly true of any release of individual-level data to other researchers, which is not covered by the table server approach. There must be strong

controls at the recipient institution to ensure data is handled properly and with due care for confidentiality (and not just security).
The answer is a graded approach: using perhaps 'quick & dirty', though effective, controls initially, relaxed as other controls are put in place. The controls can be more liberal where specific approvals are gained and public benefit proven/accepted, perhaps with specific Section 251 or other support.

5. Conclusion

Any CDW needs to have SDC controls designed in, but these need to be balanced, flexible, and use a variety of techniques. The basis of use needs to be 'anonymized data' to avoid needing to seek specific consent (see Singleton & Wadsworth 2006). However, 'anonymization' is not defined except as not 'personal data', which requires a provably intractable re-identification test to be absolute – this is unlikely to be possible for any but the most rudimentary of CDWs.

Exact re-identification can only occur if released data contains, or allows the recovery, of a unique. Even if this is not possible, it does not prevent attribute disclosure. The table server approach can be tuned to guarantee against exact identification or exact attribute disclosure, although the latter would often imply risk thresholds that would severely limit the possibilities for data release; the presence of a single zero raises the possibility of exact (negative) attribution. Nevertheless, the approach allows all recoverable 0s and 1s to be identified and the risk measures / thresholds can be tuned to provide a useful means of identifying tables that can be safely released. Its dynamic nature offers advantages over similar systems that rely on the DSO to specify the utility of data. Coupling with an analysis server provides a flexible overall system for dealing with most of the queries that might be made of a CDW. Of course, users might be able to make a case that particular data are of such high utility that they should be released even if an existing risk threshold is exceeded. These cases will always need to be dealt with on an *ad hoc* basis by the DSO.

References

Andrews, P., Sleeman, D., Statham, P., McQuatt, A., Corruble, V., Jones, P., Howells, T. and Macmillan, C. (2002) Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. J. Neurosurg. 97, 326–336

Austin, T., Lea, N., Tapuria, A., and Kalra, D. (2008). Implementation of a Query Interface for a Generic Record Server. International Journal of Medical Informatics 77(11), 754-764. ISSN: 1386-5056.

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.

Domingo-Ferrer, J. and Torra, V. (2008) A critique of *k*-anonymity and some of its enhancements. In Proceedings of ARES/PSAI, Los Alamitos, CA:IEEE Computer Society, pp.990-993

Duncan, G. T., Elliot, M. J. and Salazar-Gonzalez, J-J. (2011) Statistical Confidentiality. Springer: new York.

Eastwood, E., Magaziner, J., Wang, J., Silberzweig, S., Hannan, E., Strauss, E., Siu, A. (2002) Patients with hip fracture: subgroups and their outcomes, J. Am. Geriatr. Soc. 50, 1240–1249.

Elliot, M.J. and Dale, A. (1999) Scenarios of attack: the data intruder's perspective on statistical disclosure risk. Netherlands Official Stat. 14, pp.6–10

Elliot, M., Purdam, K. and Smith, D. Statistical disclosure control architectures for patient records in biomedical information systems. Journal of Biomedical Informatics 41 (2008) 58–64.

Fellegi, I. and Sunter, A. (1969). A theory for record linkage. Journal of the American Statistical Association 64 (328) pp.1183–1210

Grant, A., Kalra, D. and Fuller, M. (2006), Good principles and practices for a clinical data warehouse. ISO Technical Rreport 22221. Geneva: International Organisation for Standardisation.

Grant, A., Thorp, J. and Fuller, M. (2010). Deployment of a clinical data warehouse. ISO Technical Specification 29585. Geneva: International Organisation for Standardisation.

Kalra, D., Singleton, P., Milan, J., MacKay, J., Detmer, D., Rector A. and Ingram, D. (2005). Security and confidentiality approach for the Clinical E-Science Framework (CLEF) Methods of Information in Medicine 44(2), 193-197. ISSN: 0026-1270

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.

Lee, C. H., Chen, J. C., and Tseng, V. S. (2101) A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. Comput Methods Programs Biomed. 2010 May 27. [e-Pub]

Machanavajjhala, A., Gehrke, J., Kiefer, D. and Venkitasubramanian, M. (2006) L-diversity: privacy beyond k-anonymity. In Proceedings of the IEEE ICDE 2006.

Pfaff, M., Weller, K., Woetzel, D., Guthke, R., Schroeder, K., Stein, G., Pohlmeier, R. and Vienken, J. (2004) Prediction of cardiovascular risk in hemodialysis patients by data mining. Methods Inf Med. 43,106-13

Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. Journal of Business and Economic Statistics, 6(4), pp.487-500

Mokken, R.J., Kooiman, P., Pannekoek, J. and Willenborg, L.C.R.J. (1992) Disclosure risks for microdata. Statistica Neerlandica, Vol. 46. pp.49-67

Reiter, J.P. (2005) Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. J. R. Statist. Soc. A, 168, Part 1, pp.185–205

Smith, D. and Elliot, M J. (2008) A measure of disclosure risk for tables of counts. Transactions in Data Privacy. 1(1), pp.34-52

Samarati, P. and Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

Singleton, P. and Wadsworth, M. (2006) 'Practical aspects of obtaining consent for the use of personal medical data in research', British Medical Journal, Jul 2006; 333: 255 – 258.

Sparks, R., Carter, C., Donnelly, J. B., O'Keefe, C. M., Duncan, J., Keighley, T., and McAullay D. (2008) Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics Computer methods and programs in biomedicine 91(3); 208-222

Willenborg, L. C. R., and de Waal, T. (2001) Elements of Statistical Disclosure Control. Lecture Notes in Statistics, Vol. 155, Springer, New York