WP.36
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

THE REVIEW OF THE DISSEMINATION OF HEALTH STATISTICS IN ENGLAND

Supporting Paper

Prepared by Jane Longhurst, Carole Abrahams, Ann Blake, Nirupa Dattani (Office for National Statistics); Mary Grinsted (Department of Health); and Gwyneth Thomas (Welsh Assembly Government)

The Review of the Dissemination of Health Statistics in England

Jane Longhurst, Carole Abrahams, Ann Blake, Nirupa Dattani (ONS)*
Mary Grinsted (Department of Health)**
Gwyneth Thomas (Welsh Assembly Government)***

* Office for National Statistics, Segensworth Road, Fareham, UK, Jane.Longhurst@ons.gov.uk
** Department of Health, Skipton House, 80 London Road, London, Mary.Grinsted@dh.gsi.gov.uk
*** Welsh Assembly Government, Cathays Park, Cardiff, Gwyneth.Thomas@wales.gsi.gov.uk

1. Introduction

Health statistics support a wide range of work to improve and protect our health, they inform patients and the public. Many of these areas of work require detailed figures, which may raise issues about data confidentiality. Producers of health statistics must ensure that their statistics meet the needs of users while at the same time protecting confidentiality. The Review of the Dissemination of Health Statistics in England was initiated in 2005 to address disclosure issues around health statistics. The aim of the review was to produce guidance for handling health statistics in a way that ensures the public interest in the figures is met while managing data confidentiality risks.

The review was led by the Office for National Statistics (ONS) and involved representatives from the Health Departments in England, Public Health Observatories and the devolved administrations (Wales, Scotland and Northern Ireland). The approach adopted for the review was a two-stage process. The first part of the review focused on developing guidance for published tables of abortion statistics. Specific guidance for these outputs was released in July 2005, ONS (2005) and has been subsequently implemented by the Department of Health. The scope was then extended to provide more general guidance on disclosure issues for all published tables of health statistics. Throughout the development of the guidance key stakeholders were consulted via a series of workshops and in addition the guidance was released for public consultation. Following quality assurance and approval from the National Statistician and Health Minister the final guidance from the review was published in October 2006 on the National Statistics website, ONS (2006).

This paper provides an overview of the final guidance in Section 2. Section 3 describes work that is being undertaken to support the implementation of the guidance across the health domain. Sections 4, 5 and 6 detail three specific examples of practical implementation of the guidance from the Department of Health, the Welsh Assembly Government and ONS.

2. The Guidance

2.1. Scope

The principles and approach outlined in the guidance apply to all health statistics. However, the review is focused on tables derived from registration processes, administrative sources and statistical returns. It does not deal with confidentiality issues concerned with record-level information. The guidelines replace previous practices that had been adopted within the health field, such as the rule of thumb to suppress all values in tables less than 5.

The review was established specifically for published health statistics, where following release there is no control over their further use. The guidelines should be used to protect statistics released as part of a production process, however, ad-hoc releases and in particular Freedom of Information (FoI) requests are also within scope. Throughout the review it was therefore necessary to consider the implications of FoI and the guidelines were developed taking into consideration what it is or is not appropriate to withhold under the Act.

2.2. Format of the Guidance

The final guidance has been published on the National Statistics website as seven pdf documents; a main document, five working papers and a summary of the responses to the public consultation. The main document describes an approach that data providers should follow based on a general framework for addressing the question of confidentiality protection (see section 2.3). No single solution or rule is recommended instead guidance is provided, based on the steps in the framework, on how to develop solutions for different datasets. Examples are used throughout the document and more technical advice is provided in the five working papers:

1. Legal and Policy Considerations.
2. Risk Assessment
3. Risk management
4. Glossary
5. References and other Guidance

2.3. Framework for Confidentiality Protection

The guidance advises producers of statistics of six main steps to be taken in considering disclosure control in relation to tables of health data. The guidance works through each step, giving details, examples and useful references. The six steps are:

1. Determine users' requirements for the published statistics
2. Understand the key characteristics of the data
3. Are there circumstances where disclosure is likely to occur?
4. If so, would disclosure represent a breach of public trust, the law, or National Statistics policy?
5. If required, select appropriate disclosure control methods to manage this risk
6. Implement and disseminate.

2.3.1 <u>Determine users' requirements for the published statistics</u>

The first priority for producers of statistics should be that their publications meet the needs of users. It is therefore vital to identify the main users and understand why they need the figures and how they will use them. The disclosure protection used needs to have the least possible adverse impact on the usefulness of the statistics.

2.3.2 <u>Understand the key characteristics of the data</u>

It is important to have a good understanding of the data that may require protection to assess any risk of disclosure. Issues to consider include:

- the source of the data underlying the statistics
- sensitive variables
- the age of the data; older data may carry less risk of disclosure
- quality of the underlying data
- small groups of statistical units

- whether the data is event-based or residence-based

It is also important to consider the characteristics of the tables. Where tables are very simple and presented at a high level of aggregation (including geography), disclosure issues are unlikely to arise. When tables become more detailed, and the counts in individual cells are small, the risk of identification may increase and protection may be needed. If the spread of values is skewed across a table, the risk in particular cells may increase above an acceptable level. In addition, issues may arise with linked tables where risk of disclosure can increase by differencing or through combining with other data.

2.3.3 Are there circumstances where disclosure is likely to occur?
The answer to this question is the risk assessment. Risk is a function of likelihood (related to the design of the table), and impact of disclosure (related to the nature of the underlying data). Decisions on likelihood and impact should be made by those who have a detailed understanding of the statistics and experience of the interest in the figures. In order to be explicit about the disclosure risks to be managed one should consider a range of potentially disclosive situations and take action to prevent them. The situations should be used to identify those parts of the table that could lead to disclosure. Appropriate confidentiality rules should be applied to these cells. The guidance provides examples of disclosive situations but notes that it is not possible to protect against all risks and that this is a risk management rather than a risk elimination exercise.

In practice it is likely that producers of statistics will find that outputs can be placed into one of three broad risk categories and recommendations are made on the level of protection required for these three risk categories.

- Low Risk: For some health statistics the likelihood of an attempt at identification may be considered to be low if tables are disseminated at a high level of aggregation and only limited tables are produced from the one database, i.e. no risks from linking between current and future releases. A high level of aggregation reflects a reduction in disclosure risk as the size of the population of the statistic increases. Health statistics in this category will not usually require any protection beyond good table design. However, in order to prevent attribute disclosure care should be taken where rows or columns are dominated by zeros and in particular where a marginal total is a 1 or 2.
- Medium Risk: In order to ensure protection from disclosive situations for the majority of health statistics it will be sufficient to consider all cells of size 1 or 2 unsafe. Care should also be taken where a row or column is dominated by zeros.
- High Risk: For some health statistics the likelihood of an identification attempt will be higher, and the impact of any successful identification would be great, e.g. statistics on abortions, AIDS/HIV, STDs. In order to ensure protection all cells of size 1 to 4 are considered unsafe and care should be taken where a row or column is dominated by zeros. Higher levels of protection may be required for small geographical levels or for particular variables with an extremely high level of interest and impact.

The guidance outlines situations where these recommended levels of protection may need to be increased.

2.3.4 If so, would disclosure represent a breach of public trust, of the law, or of National Statistics policy?
When establishing whether confidentiality protection is required for a particular health statistic, it is necessary to consider public trust and cooperation, and legal rights and obligations, as well as national and international standards for statistics. More legal and policy considerations are provided in the relevant working paper.

2.3.5 If required, select appropriate disclosure control methods to manage this risk
An introduction is given to five main statistical disclosure control methods. Table redesign (grouping/collapsing categories, aggregating across geographies or time) is recommended as a simple method that will minimise the number of unsafe cells. If unsafe cells remain in the table, further protection methods such as rounding, cell suppression or cell perturbation (e.g. barnardisation) should be considered. If a data provider has access to the individual record level data then disclosure control methods can be implemented that adjust the data before tables are designed, e.g. record swapping. The different methods are compared and contrasted to assist the selection of the appropriate disclosure control tool. The guidance also recommends alternative methods for presenting the data, e.g. graphs, commentaries or analytical output, as an approach for providing users access to information without disclosing the underlying data.

2.3.6 Implement and disseminate.
The guidance will allow data providers to set disclosure control rules and select appropriate disclosure control methods to protect different types of health statistics that are to be published. The most important consideration is maintaining confidentiality but these decisions will also accommodate the need for clear, consistent and practical solutions that can be implemented within a reasonable time and using available resources. The methods used will balance the loss of information against the likelihood of individuals' information being disclosed.

The guidance has been developed taking into account the implications of the Freedom of Information (FoI) Act and therefore confidentiality policy developed using this guidance can be used to help decide which exemptions in the Act are relevant, and which should be cited when withholding confidential statistical information. However, FoI requests should always be considered on a case by case basis and there may be cases when decisions about a case are different to the general policy for the publication of statistics. This does not mean that the policy is wrong, since it has been developed for use in a production process. Whilst confidentiality must always be maintained, a decision made under FoI to provide information in a form different to the published outputs is compatible with this guidance.

Guidance is provided more generally on implementation, in particular the information that should be provided to users concerning the confidentiality rules and disclosure control methods.

3. Implementation

The guidance produced from The Review of the Dissemination of Health Statistics as outlined above is intended for anyone in the health community involved in the publication of health statistics. The Office for National Statistics (ONS) have given assurances that the proposals in the guidance will be implemented for their outputs. In addition the ONS will work collaboratively with the Information Centre for Health and Social Care and the Department of Health to support the implementation of the recommendations more widely.

A project within ONS has been established to coordinate this implementation, the aim being to ensure that health statistics releases throughout England are consistent with the new confidentiality guidance on disclosure and to encourage the adoption of the guidance for the other UK countries. The project board includes representatives from ONS, the Health Departments, Public Health Observatories and the devolved administrations.

Each key producer of health statistics represented on the project board is drawing up and executing plans to implement the guidance in relation to their outputs from April 2007. It is not expected that the guidance will be applied retrospectively unless there is an exceptional reason to do so. Templates (see below, 3.1) have been developed to record basic information on outputs and timetables for release and more detailed information relating to the steps within the framework for confidentiality protection and proposals for disclosure control. Where statistics are produced by organisations outside the representation of the project board, these organisations will be encouraged to follow the guidance.

The project board is responsible for coordinating these high level implementation plans and timetables. In addition the board will provide expert advice and support for implementation and facilitate consistency and sharing of best practice.

3.1 Risk assessment templates

The Welsh Assembly representative on the project board provided the risk assessment template which had been developed and was being used in Wales (see section 5). This was slightly modified by the project board members representing ONS health statistics to produce two templates, a high-level one and a low-level one.

The high-level template is an Excel spreadsheet on which the data providers can enter details of each output which it is planned to publish during the year 2007/08. Work on populating these has begun. The information entered onto this template includes, for each output:

- Source of the data, e.g. “cancer registrations”; “calculated using life tables & census data”; “smoking cessation questionnaire”.
- Output type, e.g. HSQ (Health Statistics Quarterly) article; web release; quarterly statistics.
- Name of the output
- Area coverage
- Date of publication
- Key characteristics, e.g. “contains potentially sensitive information”; “may contain low numbers”; “low sensitivity”.
- Disclosure issues and risks – High, Medium or Low
- Impact of disclosure e.g. whether it would be a breach of public trust
- Statistical disclosure controls applied – Yes or No
- Link to low level report - this is an electronic link to the corresponding low-level template

The low-level template is a Word document, and there will be one for each planned publication. This template is based around the framework described in section 2.3. The details to be entered are:

- Background to the data source
- Legal issues (collection and dissemination) - Reference to any relevant statutory arrangements relating to the data, e.g. population statistics act
- Key characteristics of the output - List of sensitive variables, age and quality of data, coverage, population base, possibility of linking to other published tables
- Evidence of risk in the output - A summary of the risk assessment including:
  - a note of disclosure scenarios considered
  - potentially unsafe cells
  - arrangements with data provider (e.g. if data are provided by another government department). Colleagues' views should also be recorded here.
- Proposals for mitigating risk in publishing output, including list of options
- Conclusions and details of disclosure control methods to be used: Which option was chosen and why, and description of methods to be used.
- Review process

The next three sections of the paper provide practical examples of how the guidance is being implemented for different outputs.

4. Abortion Statistics

The Department of Health (DH) undertakes the statistical processing and analyses of notifications of abortions for England and Wales. This includes the release and publication of statistics derived from the information contained within the notification. Abortion data is very sensitive and therefore the impact of any identification or disclosure from these statistics is considered to be high. Abortion data attract a lot of attention from the media, MPs and Peers, the public and lobby groups and is likely to be scrutinised closely, particularly at its margins.

DH receive a lot of requests for abortion data some of which are potentially disclosive, e.g. numbers to girls aged 11, 12 & 13 years old, medical conditions of late abortions. From experience of high profile cases in 2001/02 it was known that statistics like these, possibly used with other information, could be used to identify and target individuals. As a result 2002 data due to be published in 2003 was held back and only a skeleton publication was released. This limited publication and refusals to release requested information, resulted in DH being seen as overly cautious and accused of hiding information. Clearly guidance was needed in interpreting the National Statistics Code of Practice in order to balance the data confidentiality risks with the public interest. The first part of the Review of the Dissemination of Health Statistics focused on abortion statistics and attempted to address these concerns.

The guidance provided details on how to identify cells within tabulated statistics where the risks of a breach of confidentiality were unacceptable ("unsafe cells"). The risks within the publication were identified and were reduced largely by redesigning tables. However, where table redesign proved to be impossible then suppression was applied to cells with fewer than 5 cases at National level or fewer than 10 cases at sub-national level and to highly sensitive variables such as gestation weeks in tables of terminations by medical grounds. The same principles were also applied to tables showing rates and percentages and to ad hoc requests for data.

The guidance is followed for all requests and for the vast majority it is unquestionably useful. It is especially useful to be able to point questioning customers towards the protocol on the internet and for them to know it is a health statistics wide issue rather than a data provider's decision. However, in a very few instances the information suppressed does seem overly cautious, more so to the customer than to the data provider, who understands that the rules work for the majority of cases.

4.1. Example - Abortions performed by the British Pregnancy Advisory Service

In order to illustrate the disclosure issues related to abortion statistics the following example of a Parliamentary Question is described. Through contractual arrangements with Primary Care Organisations, some approved independent sector places of termination perform NHS-funded abortions. The query related to the number of early medical abortions performed by the British Pregnancy Advisory Service (BPAS) at the request of the National Health Service (NHS) in each of the last five years, broken down by (a) age of the woman, (b) gestation of the pregnancy and (c) region.

There are three key items relating to an individual abortion which need to be protected. These are: the details of the woman whose pregnancy was terminated, the identity of the practitioner who carried out the termination, and the identity of the hospital or clinic where the abortion took place.

Therefore in answering the query the following were considered:

- Confidentiality of the patient - control for small numbers, e.g. counts less than 10.
- Confidentiality of the doctor - check original documents for doctors' names to make sure there were at least three doctors performing terminations at any one clinic.
- Confidentiality of clinic - check that there was more than one BPAS clinic per region.
- Confidentiality of the data - make sure no similar data had been previously published that could be used deliberately or inadvertently to disclose small numbers.

An extract from one example table from the release is shown below, suppressed cells are indicated by '..':

Table of counts of abortions performed under 8 weeks at BPAS clinics

Region	2002	2003	2004
East	17	43	..
East Midlands	..	..	78
London	249	362	582
North West	52	74	..
North East	..	..	97
South East	409	507	913
South West	44	85	192
West Midlands	376	348	597
Yorkshire and Humber	83	173	171
Total	**1252**	**1636**	**2683**

The data was grouped within the tables in various ways in order to maximise the amount of information provided and minimise the number of cells that had to be suppressed. The data was compared with similar data extracts for clinics run by other agencies in case an equivalent request was made for these other agencies. Data also had to be checked for disclosive cells in groups that could be derived from other published data, e.g. privately funded abortions, NHS hospital terminations within the same regions and more than 9 weeks' gestation. In some cases small counts were suppressed and in order to protect these counts further, secondary suppressions of counts greater than 10 may have also been suppressed. Care will need to be taken with future releases to ensure that these counts and other totals are not revealed, otherwise the protection for this release could be compromised.

5. Implementation of the Guidance by the Welsh Assembly Government

During the time when the Health Statistics confidentiality guidelines were being prepared, the Statistical Directorate of the Welsh Assembly Government was also considering statistical disclosure issues in health statistics and across the range of subject areas covered in the Directorate. Internal guidance has been written and is being used, based largely on the Health Statistics guidance but widened in scope to cover all statistical areas dealt with by the Directorate.

A key part of the guidance is the process of assessing and documenting risk and, as an aid to this process, a risk assessment template has been devised in consultation with staff. It is felt to be important for two main reasons. Firstly so that consistent decisions about disclosure control can be taken for each dataset in advance of the receipt of ad-hoc requests; and secondly, so that there is documentary evidence that disclosure risk has been considered. The risk assessment details evidence of risk in the dataset based on the kind of issues described in the Health Statistics guidance, such as, sensitivity, geography and denominators, low cell counts, zero value cells, age and quality of the data and so on. Users of the template are encouraged to seek the views of colleagues working in a practical way in the subject area, look for examples of disclosure control in similar or related datasets and to think about scenarios where disclosure might occur. Options for mitigating risk are considered and conclusions drawn about future practice and methodology for disclosure control if necessary. Because they may detail intended methods for disclosure control, risk assessments are intended to be internal documents only. As a public acknowledgement of the issue the

Directorate is planning to add a standard phrase to publications stating that statistical disclosure risk has been considered, risk is felt to be low/medium/high and as a result data has been modified/not modified.

The template is being gradually introduced for all datasets in the Statistical Directorate. It has been employed so far in such diverse areas as health statistics and economic statistics. The format has worked best where the responsible statistician is also the data collector, where it provides a useful aide memoir to the thought process of considering the risk of disclosure. It is less useful where data dissemination is already governed by the rules of a third party data provider. The guidance used in the Directorate together with the risk assessment template will be subject to a review during the next few months. For health statistics it is working well and following a workshop on disclosure control organised by the Welsh National Public Health Service and the Welsh Health Analysts' Network, partner organisations in Wales have expressed an interest in utilising it as a basis for their own statistical disclosure control decisions.

5.1 Example - Community Contraception Statistics

As an example of how the risk assessment template has been used, the publication of community contraception statistics is considered. Data is collected annually from NHS Trusts in an electronic form. The dataset is aggregate data and relates to the numbers of first face-to-face contacts of patients with the service in the financial year by age (individual year of age for women aged less than 20), reasons for attending the clinic and method of contraception chosen.

The dataset contains items which may be sensitive in some cases, for example, the contraception methods chosen by young girls. Low cells and real zeros may be disclosive where many dimensions are tabulated but disclosure is unlikely otherwise. It is possible that potential intruders have access to local information which would help them find, say, individuals who have chosen unusual contraception methods given their age but it is not clear what that information might be since clinics would not break their patients' confidentiality.

A degree of uncertainty is present in this dataset since the coverage is only a proportion of all contraception consultations; it only relates to community clinics and excludes consultations with GPs and private clinics. Also, the data relates only to first contacts in the financial year which again means that not all visits are included introducing further uncertainty. The lowest level of disaggregation is NHS Trusts and there is no possibility of differencing.

After consideration and discussion with NHS professionals, it was felt that where two dimensions are involved e.g. method and age or NHS Trust and age, risk is fairly low but where all three are involved the risk increases with a determined intruder. Thus it was concluded that in routine and ad-hoc tabulations only 2 of the 3 possible dimensions would be used. Trusted users might be allowed access to more detailed data but only having signed a confidentiality agreement.

6. Disclosure control for health data with rare events: Development of a method for presenting conceptions to girls aged under 18 by small area

As outlined in Section 3 the ONS will be reviewing all outputs of health data as part of the implementation of this guidance. One area where work has already been carried out is for conception statistics. In order to meet a requirement for information to support action at local level towards achieving the Public Service Agreement target to reduce the under 18 conception rate by 50 per cent by 2010 a project was undertaken to produce an output presenting under 18 conceptions by ward.

There are around 40,000 conceptions to girls aged under 18 in England and Wales each year. But the number of cases at ward level is relatively low. Therefore a method of presenting these data has been developed which will provide useful information by small area whilst protecting confidentiality of individuals.

Due to the sensitive nature of the data, the small area of the geography, and the focused nature of the age group of interest the methodology combines three statistical disclosure control methods.

First the data for three years were combined. This served two purposes:

- It smoothed the data and thereby reduced the impact that natural variation in rare events can have on understanding trends over time
- It increased the ambiguity of small numbers so that it is not possible for a potential intruder to identify individuals

The data for England and Wales were then divided into five bands (quintiles). Wards with the lowest rates were allocated to quintile 1 and those with the highest rates to quintile 5. Data for wards with a population of fewer than 30 girls in the 15 to 17 age group were then suppressed. The upper and lower limits for each band were examined to assess whether publication of these, in combination with the availability of population estimates for the group of interest, may allow an intruder to unpick the data and calculate actual numbers of teenage conceptions in a particular ward. It was established that this may be a possibility, and therefore only the lower limit for the wards with the highest rates will be made available.

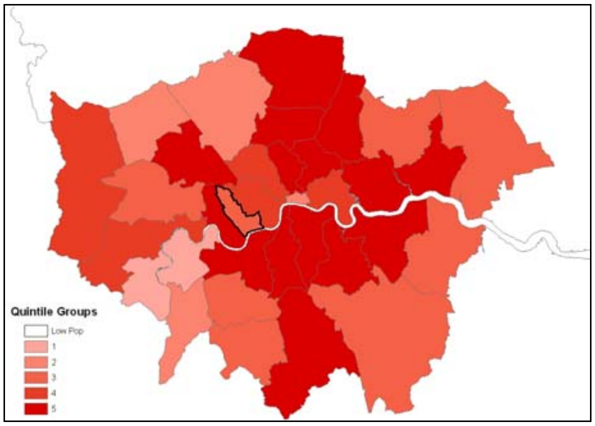
Users would be able to identify which quintile a particular ward falls into, and also make a comparison between wards in the country using the Neighbourhood Statistics mapping functionality on the ONS website. In this way the under 18 conceptions rates for individual wards are not published, but because they can see the geographic distribution of the data, users are still able to identify areas with high rates that they may wish to target. Further information on these data can be obtained from ONS (2007).

Using Kensington and Chelsea the following example shows how the ward and local authority (LA) level maps provide extra information about under 18 conception rates within a given area.

The first map provides a view of LAs within the London Government Office Region for 2001 - 2003. The map allows comparisons of rates between LAs and also shows

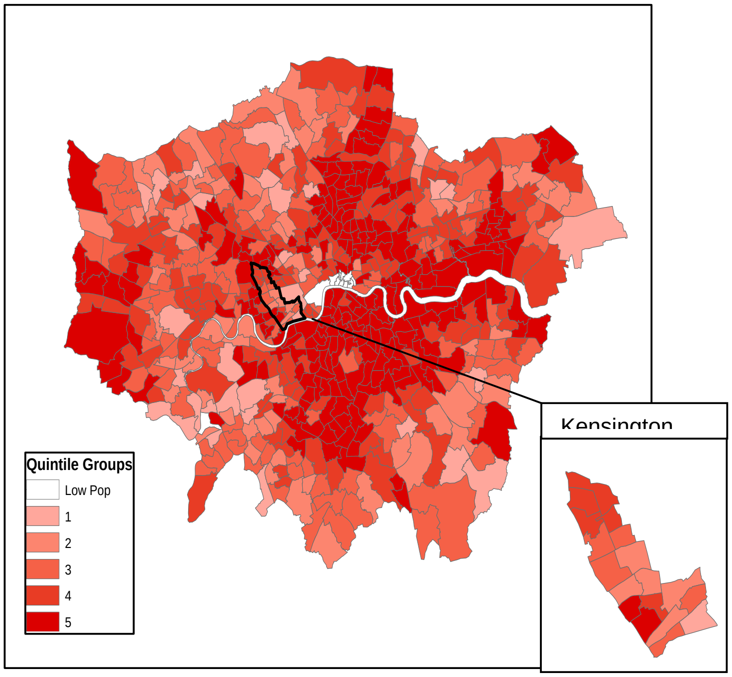
LAs with high rates of under 18 conceptions, i.e. those in quintile group 5. For Kensington and Chelsea (as highlighted) we can see that this LA is in quintile 3. Therefore the rate for this LA is in the middle quintile.

Under 18 conceptions data for Local Authorities in London, Jan 2001 – Dec 2003



The map below shows the information at ward level for the London region, for 2001-2003. This shows that there is significant variation in under 18 conception rates by ward within some local authorities. We can now see that there are also wards within Kensington and Chelsea with the highest rates of under 18 conceptions.

Under 18 conceptions data for wards in London, Jan 2001 – Dec 2003



7. Conclusion

In using and releasing health statistics there is a risk, generally with small numbers, of identifying individuals. To address this, the Department of Health in England asked the National Statistician to provide it with guidelines for disseminating health statistics, in a way that balances data confidentiality risks with the public interest in the use of the figures. Guidance has been developed based on a framework for addressing issues concerning confidentiality. No single solution or rule is recommended; instead data providers are encouraged to develop solutions for different sets of statistics based on the steps in the framework.

Work is being undertaken to implement the guidance. In particular the Welsh Assembly Government has developed templates to aid in the risk assessment process and to document the decisions made. The Department of Health are using specific guidance developed for abortion statistics for all annual and ad-hoc releases. The ONS has also developed new ways to release conception statistics without disclosing the underlying small counts. These are just three specific examples. Work on implementation is being coordinated by the Implementation Project Board to ensure consistency and encourage best practice.

References

ONS (2005) Disclosure Review for Health Statistics, 1st report, guidance for abortion statistics,
http://www.statistics.gov.uk/downloads/theme_health/abortion_stag_final.pdf)
ONS (2006) Review of the Dissemination of Health Statistics: Confidentiality Guidance, http://www.statistics.gov.uk/about/Consultations/disclosure.asp.
ONS (2007) Conceptions, Under 18's: Local Analysis
http://neighbourhood.statistics.gov.uk/dissemination/Info.do?page=news/newsitems/6-august-2007-conceptions---under-18s-local-analysis.htm