

Disclosure Review for Health Statistics
1st Report - Guidance for Abortion Statistics

July 2005

Contents

	Page
Summary	1
Conclusions	3
1. Introduction	5
2. Key Aspects of Abortion Statistics	7
3. Why is Confidentiality Protection needed?	10
4. Risk Management	16
5. Disclosure Control Methods	27
6. Findings	32
Appendix 1 – List of Tables	34
Appendix 2 – Conceptions Statistics	35
Appendix 3 – Disclosure Control Methods	38

Summary

1. The National Statistics Code of Practice and underpinning Protocol on Data Access and Confidentiality provide principles for the protection of the confidentiality of National Statistics. A need for more detailed guidelines on how to interpret these policy statements in practice when using and releasing health statistics has been recognised. An Office for National Statistics led project has been established to undertake a review of disclosure issues around health statistics and to produce guidelines for handling health statistics across the health community, in a way that balances data confidentiality risks with the public interest in the use of the figures.
2. This report is the first deliverable for the review and provides standards and guidance for protecting the confidentiality of abortion statistics. The guidance provided is based on best practice as of the date it is produced. A second project deliverable will document the process undertaken to reach the design standard for this case study and will be extended to cover all other health statistics within scope.
3. This report covers the legal, policy and ethical frameworks underlying the confidentiality protection measures needed for abortion statistics. This guidance focuses on the disclosure issues associated with tabular publications of abortion statistics. It is not within scope to provide guidance for the confidentiality issues associated with individual level data used for administrative or research purposes. It presents a number of scenarios where a breach of confidentiality might arise. For each such scenario it identifies the particular parts of the abortion statistics that pose a risk of a breach of confidentiality – the cells within the tabulated statistics where the risks are unacceptable. These cells are described as ‘unsafe cells’.

Identifying unsafe cells

4. From consideration of the risk scenarios the unsafe cells for abortion statistics are those for which
 - the count is zero unless no other value is logically possible, or
 - the count is below 5 for a Government Office Region in England, the country of Wales or a larger area, or
 - it is less than 10 and the area concerned is smaller than Government Office Region in England or the country of Wales, or
 - it is less than 10 and the variables are considered highly sensitive, or
 - the count is associated with at most 2 practitioners, or
 - it is associated with at most 2 hospitals.
5. The variables that are considered highly sensitive are:
 - Young ages (<15)
 - Late gestation (over 24 weeks)
 - Procedure by gestation
 - Medical conditions
6. Simple calculations such as rates or percentages do not necessarily make an unsafe cell safe and should not be used to protect data unless it can be demonstrated that one cannot work back to the original count. For abortion statistics rates or percentages should only be calculated from safe cells.

Mitigating the disclosure risks in unsafe cells

7. Once identified, the risk of disclosure will need to be reduced. Methods for managing this risk include:

- table redesign
 - using area of residence rather than place of termination in constructing statistics
 - suppression
 - restricting geographies and categories
 - aggregating data over several years
8. Each method has advantages and limitations and these are described in the report. It is for those within the Health community to use professional judgement in determining which combination of method is appropriate to each situation where a confidentiality risk is identified.
9. The conclusions are made for the release of abortion statistics and are particular to the risks, utility and design of the data.
10. The guidance is written for those within the health community who are involved in the release and publication of abortion statistics. The guidelines will be mandatory for the Office for National Statistics. The intention is that they will be adopted by the Department of Health and The Health and Social Care Information Centre. It is hoped that others within the health community will also adopt the guidelines. The conclusions should be implemented for all published outputs of abortion data e.g. in cases where the data published can be used to make direct inferences about abortions.
11. The report considers how the conclusions for this review will affect the statistical disclosure control requirements for conceptions statistics. The requirements will not be affected for numbers and rates of conceptions which should continue to be published using the current disclosure control methods. For percentage conceptions leading to abortions and rates of conceptions leading to abortions additional disclosure control will be required.

Conclusions

The conclusions are listed here in the order that they appear in the report. These conclusions are made for the release of abortion statistics and are particular to the risks, utility and design of the data.

Risk Management (Chapter 4)

- 1 Within a release of abortion statistics unsafe cells are defined as being counts of abortions that are:
 - zero unless no other value is logically possible
 - less than 5 for Government Office Region in England, the country of Wales or any larger geography
 - less than 10 for any geography below the Government Office Region in England or the country of Wales
 - less than 10 for highly sensitive variables
 - associated with either 1 or 2 practitioners
 - associated with either 1 or 2 hospitals

The variables that are considered highly sensitive are:

- Young ages (<15)
 - Late gestation (over 24 weeks)
 - Procedure by gestation
 - Medical conditions
- 2 Simple calculations such as rates or percentages do not necessarily make an unsafe cell safe and should not be used to protect data unless it can be demonstrated that one cannot work back to the original count. In order to keep statistical disclosure control rules clear and consistent for abortion statistics rates or percentages should only be calculated from safe cells.

Disclosure Control Methods (Chapter 5)

- 3 In order to ensure that unsafe cells in the abortion statistics are disguised table design should be used as a preliminary protection method. Redesign should be implemented taking into account the information required by the main users of the data.
- 4 Statistics should only be constructed using area of residence rather than place of termination thus reducing the risk of disclosure from counts of events that are associated with 1 or 2 practitioners/hospitals.
- 5 If unsafe cells exist in tables after redesign these should be removed using suppression methods (primary and secondary).
- 6 In order to avoid resource intensive analysis of disclosure by differencing the abortions data should not in general be published on geographies that are non-coterminous with Strategic Health Authorities (SHAs) or Primary Care Organisations (PCOs) in England or regions or Local Health Board (LHBs) in Wales or for non-standard variable categories.
- 7 In order to release more detail some data should be published aggregated over a number of years. To keep the methods for disclosure control clear, consistent and easy to implement the conclusion is made that rolling aggregates are not produced but years are aggregated independently.

Implementation (Chapter 6)

- 8 The guidelines should be implemented for all published outputs of abortion data e.g. in cases where the data published can be used to make direct inferences about abortions. The guidelines should be implemented as soon as possible. The Department of Health and the Office for National Statistics should work together in order to implement these guidelines for the annual bulletin release for 2003 and 2004.
- 9 Users should be made aware of what constitutes an unsafe cell within the abortion statistics. The user should also be told that the method used to protect the table is predominantly table redesign used to minimise the number of unsafe cells that require suppression. The impact on the quality of the data will be that in some cases less detail will be displayed and suppressions will mean that some information is removed from the table.

1. Introduction

Background

1.1 The National Statistics Code of Practice and underpinning Protocol on Data Access and Confidentiality give a guarantee that producers of National Statistics will not put into the public domain any statistics that are likely to identify an individual unless specifically agreed with them. The Code of Practice and Protocol provide principles on how to ensure that this confidentiality guarantee is met. A need for detailed guidelines on how to interpret these policy statements in practice when using and releasing health statistics has been recognised. The particular concern for a producer of National Statistics is how to achieve a balance between the utility of health statistics and the confidentiality of informants and other individuals.

1.2 The following was announced on the Department of Health website in February 2005:

In using and releasing health statistics there is a risk, generally with small numbers, of identifying individuals. There is therefore a need to review how we release our statistics to minimise risks of disclosure of individuals and also to ensure that adequate information is available to users to meet requirements for health purposes.

To address this, the National Statistician, Len Cook, has agreed to undertake a review of disclosure issues around health statistics to report by summer 2005. The terms of reference will be:

"To provide the Department of Health and the new Health and Social Care Information Centre with guidelines for interpreting the National Statistics Code of Practice and associated protocols in the handling of health statistics across the health community, in a way that balances data confidentiality risks with the public interest in the use of the figures."

1.3 An ONS led project was established to undertake this review and produce the guidelines described above.

1.4 The objectives of the review are as follows:

- To ensure that the guidelines for interpreting the National Statistics Code of Practice promote clear and consistent statistical disclosure control practice for handling health statistics that balance the risks of identification with the needs and responsibilities of users. The guidelines will be mandatory for the Office for National Statistics. The intention is that they will be adopted by the Department of Health and The Health and Social Care Information Centre. It is hoped that others within the health community will also adopt the guidelines.
- The guidelines should be written in such a way that basic principles can be understood and applied consistently across producers of official statistics and the health community.
- To provide producers of health statistics with clear and relevant advice on how to implement the conclusions of the review and how to handle issues that arise, in particular when serving users.

1.5 The scope of the review includes the consideration of:

- statistical disclosure control practices for health statistics derived from register, survey and administrative sources.
- disclosure control practices for any micro-data that are to be published or made readily available under license.
- the identification (indirectly or directly) of health professionals.
- clear rules and policies that are compatible between different agencies.

- inconsistencies that currently exist.
 - guidance on data sharing, particularly of confidential micro-data, where the sharing is for the publication of statistics.
 - ensure that the conclusions of the review are consistent with the Freedom of Information Act (2000) and other responsibilities that already exist, e.g. Data Protection Act (1998), Human Rights Act (1998), Health and Social Care Act (2001).
- 1.6 It will not be within the scope of the review to consider:
- the administrative use of confidential or anonymised data or micro-data not used for the derivation of statistics.
 - guidance for the provision or transfer of administrative data within the health community.
 - tools for protecting confidentiality.
- 1.7 The reporting process for this review is in two parts. The first part of the review will focus on developing guidelines for interpreting the Code of Practice and associated protocols to arrive at a statement of standards and guidance for the abortion statistics that are due for publication by the end of July. The second part documents the process undertaken to reach the design standard for this case study and will be extended to cover all other health statistics within scope. The aim is to provide final guidelines that are practical, clear and comprehensive.
- 1.8 Two reports will be provided:
- 1st Report - Guidance for Abortion Statistics
 - National Statistician's Review of Disclosure in Health Statistics - generic standards and guidance with interpretation of the Code of Practice and Protocols for health statistics
- 1.9 This report covers the first stage of the review providing standards and guidance for protecting the confidentiality of abortion statistics. The guidance is written for those within the health community who are involved in the release and publication of abortion statistics. The guidelines will be mandatory for the Office for National Statistics. The intention is that they will be adopted by the Department of Health and The Health and Social Care Information Centre. It is hoped that others within the health community will also adopt the guidelines. The guidance provided is based on current best practice as of the date it is produced.
- 1.10 The next Chapter of this report provides details of the abortion statistics to which these standards and guidance should be applied. Chapter 3 discusses the policy, legal and ethical issues surrounding the need to protect the confidentiality of these data. Chapter 4 and 5 describe how to assess the risk of the data set and the methods that can be used to reduce the risk to an acceptable level. The final Chapter of the report summarises the findings and provides conclusions on the implementation of the guidance.

2. Key Aspects of Abortion Statistics

Source of Data

- 2.1 The Abortion Act 1967 permits termination of a pregnancy by a registered medical practitioner subject to certain conditions. In its application to England and Wales, regulations made under the Act require any such termination to be notified within fourteen days to the Chief Medical Officer of the Department of Health (DH) or to the Chief Medical Officer of the National Assembly of Wales according to where the termination takes place. The DH undertakes the statistical processing and analyses of the notifications for England and Wales. This includes the release and publication of statistics derived from the information contained within the notifications.
- 2.2 The information contained in the abortion notification form covers the following:
- Details of practitioner terminating the pregnancy (name, address, GMC number)
 - Certification including the name and address of the two doctors who provide the opinion to say that the woman has grounds for an abortion
 - Patient's details (date of birth, postcode, ethnicity, marital status, hospital/clinic number or NHS number, parity – number of previous pregnancies resulting in (i) live or still birth, (ii) spontaneous miscarriage or ectopic pregnancies, or (iii) an abortion)
 - Treatment details (name and place of termination, funding (e.g. NHS, NHS agency or Non-NHS), feticide, surgical terminations, medical terminations)
 - Gestation
 - Grounds
 - Selective termination
 - Chlamydia screening
 - Complications
 - Death of woman
- 2.3 Abortion data should be considered as very sensitive and therefore the impact of any identification or disclosure from these statistics is considered to be high. Information on younger women, late gestations and certain medical conditions are particularly sensitive. 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.

Outputs/Publications

- 2.4 The responsibility for disseminating abortion statistics is held by the Department of Health and this department currently publishes the statistics derived from the information contained within the notifications. Prior to 2002 the statistics were published by the ONS on behalf of the DH. The main outputs from the DH are annual bulletins. The most recent statistical bulletin covering abortion data for 2003 was published in August 2004 on the DH website. In the introduction of the bulletin it states that 'the format of the tables presented in this bulletin has been changed to reflect concerns over issues of privacy and confidentiality'. Subject to the outcome of this review and following the standards and guidance produced, the DH plans to issue more detailed data for 2003 alongside data for 2004.
- 2.5 The proposed annual bulletin provides abortion data by place of residence for England and Wales and some information at the Strategic Health Authority (SHA) in England or regions in Wales and Primary Care Organisation (PCO) level in England and Local Health Board (LHB) in Wales. The tables provide the abortion data broken down by a range of variables: age, marital status, purchaser, statutory grounds, procedure, ethnicity, parity, complications, gestation period, medical conditions, non-residents. The data are displayed as counts, percentages or rates. A detailed list of the tables proposed for publication in the annual bulletin is provided in Appendix 1.

- 2.6 The DH also provides tables of statistics not contained within the bulletin as ad-hoc releases in response to requests from the public for information on abortions. The National Assembly for Wales also disseminates information on abortions relating to Welsh residents in a statistical bulletin and in response to ad-hoc requests. High level data on abortion statistics are also published in the ONS's Health Statistics Quarterly (HSQ) and the ONS use the data to compile statistics on conceptions.
- 2.7 Under the Abortion Regulations, information sent to the Chief Medical Officer (CMO) about an abortion can only be disclosed to the persons and for the purposes set out in the Regulations. Restricted notice information can be disclosed: to authorised staff of the DH, National Assembly for Wales and the ONS. Restricted notice information can be disclosed to others for the purposes of bona fide scientific research (where agreed by the CMO and released under a confidentiality agreement); and also in other limited and specified circumstances.
- 2.8 The statutory prohibition on disclosure for restricted notice information requires that only data that are sufficiently abstract from the information notified to the CMO can be published. Legal opinion has been obtained to the effect that data will not be sufficiently abstract where the numbers of cases involved are small and/or where there is a risk of identification of the individuals involved by virtue of the fact that the information disclosed could be put together with other information which is, or may become, available.
- 2.9 It is not within the scope of this 1st Report to provide guidance for the provision or transference of data within the health community for administrative purposes nor to consider disclosure control practices for identified micro-data not involved in the publication of statistics. This 1st Report will not review issues associated with the secure transfer of individual level micro-data on abortions either to ONS or scientific researchers. This guidance will focus on the disclosure issues associated with publications of abortion statistics, specifically those that are published in the annual bulletin, any information released by DH as an ad-hoc request or released by those who have access to micro-data under the Regulations. The conclusions of the review may have an impact on the confidentiality protection required for statistics on conceptions in cases where the data published can be used to make direct inferences about abortions.
- 2.10 There are a number of users and uses of the abortion data and these are detailed in Chapter 4.

Current Disclosure Control Practice

- 2.11 For 2003 data the DH used table redesign, suppression and percentages/rates to protect the confidentiality of published data on abortions. The current guidance is that:
- no cell can be published if the true count in the cell is less than 10
 - zeros can be published
 - percentages and rates are rounded
 - rates can be published if the denominator is greater than 1000.
- 2.12 These rules were based on the legal and technical expertise that was available at the time and within the period given for preparation and publication. For some variables at a high geographical level it might be argued that the threshold of 10 is high, since it does not expressly relate to guidance provided in the Code of Practice or Protocol or a specific intruder scenario. This may lead to higher levels of suppression than are necessary to maintain the confidentiality of the restricted notice information. However, the sensitivity of these statistics needs to be considered.

- 2.13 Not protecting zeros in outputs could lead to a disclosure risk in certain circumstances. When publishing percentages or rates an intruder should not be able to return to an original count by multiplying by published denominators either available within the release or from external data sources. If an intruder can do this then care needs to be taken to ensure that the original counts are not disclosive. Suppressing rates where the denominator is less than 1000 might again be thought of as high and is not expressly related to policy or an intruder scenario. In some instances this rule of suppression based on the denominator rather than the count leads to inconsistencies in disclosure protection.

3. Why is Confidentiality Protection needed?

- 3.1 This Chapter of the report details why we need to protect the confidentiality of the abortion statistics. Firstly the policy framework for confidentiality of Official Statistics is outlined as well as how this policy framework should be applied to protecting abortion statistics. The legal and ethical requirements for protecting confidentiality are also provided.

Requirements in the policy framework for Official Statistics

- 3.2 The '*Fundamental Principles for Official Statistics*' were adopted during the 47th session of the UN Economic Commission, April 1992. Principles of particular relevance for this review include:

Principle 1 - "Official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by statistical agencies to honour citizens' entitlement to public information."

Principle 2 - "To retain trust in official statistics, statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data."

Principle 6 - "Individual data collected by statistical agencies for statistical compilation...are to be strictly confidential and used exclusively for statistical purposes."

Principle 7 - "The laws, regulations and measures under which the statistical systems operate are to be made public."

Principle 8 - "Co-ordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system."

- 3.3 This review has been conducted in a manner consistent with the UN principles. Using abortion statistics as the exemplar, the review has addressed:

- The public dissemination of health statistics of *practical utility*,
- The *professional considerations* of disclosure control methodology and output design for health statistics, taking into account users' needs,
- The *confidentiality* of the underlying health data, and its further statistical uses, and,
- The *legal and professional framework* for health statistics.

- 3.4 The *UN Fundamental Principles* establish the requirement for practical utility of official statistics, and the confidentiality of the data from which they are derived. The *National Statistics Code of Practice* provides guidance as to how the balance between these principles should be achieved.

- 3.5 The National Statistics Code of Practice was published in 2002. Principles in the Code of particular relevance for this review include:

1(a). Relevance - "National Statistics will meet the needs of government, business and the community, within available resources."

1(b). Relevance - "Users' views are essential in ensuring the relevance of National Statistics. Development and implementation of policy and programmes will be based on effective consultation."

2(c). *Integrity* - "All methods used to produce National Statistics, and the reasons for their use, will be publicly available, including any managerial direction impinging on professional conclusions."

3(d). *Quality* - "Producers of National Statistics will support the development and use of standard practices."

5(a). *Protecting Confidentiality* - "The National Statistician will set standards for protecting confidentiality, including a guarantee that no statistics will be produced that are likely to identify an individual unless specifically agreed with them."

- 3.6 It should be noted that in its principles, the *Code* introduces some qualifications to the equivalent UN Principles. User's needs are to be met 'within available resources'. No statistics will be produced that 'are likely to identify' an individual. It is within these qualifications that balances between principles can be achieved. If identification of any statistical unit is to be *absolutely* prohibited in *all* possible circumstances, very few statistics would ever be published. Similarly, if statistics agencies attempted to meet all the needs of all users, the demand on resources would be unmanageable. The Protocols that underpin the *Code* assist producers of statistics to balance *Code* principles through detailing some of their necessary qualifications.
- 3.7 The *Protocol for Data Access and Confidentiality* underpins the *Code's* confidentiality principle, and provides some guidance to the necessary qualifications within it. The Protocol sets out the standards for maintaining the Confidentiality Guarantee in the *Code*. Part 1 of the Protocol states:
- "Statistical disclosure control methods may modify the data or the design of the statistics, or a combination of both. They will be judged sufficient when the guarantee of confidentiality can be maintained, taking account of information likely to be available to third parties, either from other sources or as previously released National Statistics outputs, against the following standard: It would take a disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit to others, or to reveal information about that unit not already in the public domain."*
- 3.8 The terms that add qualification to the 'likely to' phrase in the confidentiality guarantee are: '*intruder*', and '*disproportionate time, effort and expertise*'.
- 3.9 Other than to distinguish one unit from another for statistical purposes, the statistician or researcher has no interest per se in the individual statistical unit. In contrast, an intruder is someone who for whatever reason wishes to distinguish one statistical unit in order to treat that unit separately and/or differently from the other statistical units in the dataset, for a non-statistical purpose. For the purposes of determining the levels of protection for official statistics, the threats presented by intruders can be grouped into a number of types - known as 'intruder scenarios'.
- 3.10 The intruder scenarios for different statistics can vary enormously, and the Protocol can in no way address them all. Therefore the consideration of relevant 'intruder' scenarios and the 'time, effort and expertise' that may be employed by intruders to undo the confidentiality protection of a particular statistic is a responsibility of the authority that publishes the statistic. This guidance constitutes advice to the DH and other publishers of health statistics as to what might constitute an appropriate consideration of these factors. For abortions statistics, the intruder scenarios are detailed in Chapter 4.
- 3.11 Other key terms used in the Protocol are 'identify', and 'to reveal information...not already in the public domain'.
- 3.12 The term 'identify' is used frequently in legislation. For example, to distinguish personal census information from census information, the Census Act (1920 as amended) states

"personal census information' means any census information which relates to an identifiable person or household." The term 'identify' is usually reserved in legislation for the action of recognising or selecting by analysis the characteristics of a particular person or thing.

- 3.13 The Census Act makes it an offence to disclose any information that relates to an identifiable person or household to another person, without lawful authority. Note the term 'disclose' is usually reserved in legislation for the action of transferring information (identifiable or otherwise) from one party to another.
- 3.14 Thus the intention of the phrase in the protocol "*...identify a statistical unit to others, or to reveal information about that unit not already in the public domain*" is to require that official statistics do not allow for an intruder to select or recognise by analysis the characteristics of a particular individual statistical unit, such that the identity of the unit could be disclosed to others with confidence as to the correctness of the selection and recognition. The statistic must allow for other information being available to the intruder in addition to this statistic.
- 3.15 Identification of an individual will disclose one or more attributes of that individual. The attributes disclosed are those of the table containing the information, and usually the fact of presence in the UK at the time of data capture. This latter may not, in most circumstances, be important, but it is worth noting that Solicitors and even the Courts have sought this information from publishers of statistics to gain proof of presence in the UK for the purposes of serving divorce petitions, residency claims, etc.
- 3.16 The phrase also requires that where an individual statistical unit is not uniquely identified, but where an intruder can confidently assign attributes to a particular unit that are likely to relate only to that unit, the intruder should not be able to reveal information about that statistical unit from the statistic that is not otherwise in the public domain.
- 3.17 For example, a statistic might show that a certain ward has one police officer. The Confidentiality Guarantee is breached if that statistic allows for the recognition or selection by analysis of the characteristics that distinguish that police officer from all other statistical units to the extent that the intruder can become reasonably confident of the police officer's unique identity.
- 3.18 If the intruder does not discover the unique identity of the police officer, but does manage to reveal information about the police officer that is not in the public domain, then again the Guarantee is breached. Thus it is a breach for an intruder to be able to say "I don't know the police officer's name, but I do now know he is male, 50, and a divorcee."

The Issue of 'Self-identification'

- 3.19 An individual that can recall their circumstances at the time of data collection (whether survey, census, or administrative collection) will be able to work out of which cell in any published table their information forms part. They will have 'identified themselves', but they will have done so because they know what attributes were provided in the data collection, and they know other relevant information about themselves to assist in the selection by analysis of the cell they are contributing to. Where this cell has a population greater than 1, this generally is not a significant matter for the designers of official statistics. Where the cell of self-identification has a population of just 1, or through subtraction or deduction using other available information becomes in effect a population of 1, then this becomes a matter which the Protocol requires to be taken into account. This circumstance is the self identification of uniqueness - the discovery by an individual of their uniqueness in the population of the statistic, which is something they might not have known about themselves before. This can lead to potential harm or

distress for the individual, or may cause the individual to claim that the official statistics are inadequate to protect them, and therefore others.

- 3.20 Therefore the Data Access and Confidentiality Protocol requires that publishers of National Statistics are required to consider the issue of self identification, and in particular to consider and, if necessary, take into account before publication:
- The lawfulness and ethics of publishing statistics that allow self-identification,
 - The threat to the credibility of the confidentiality guarantee,
 - That self-identification will not lead, by subtraction or deduction, to the identification of others,
 - That the self-identification will not cause the individual substantial damage or substantial distress.

Sensitivity of information

3.21 The Data Protection Act recognises that some personal information is 'sensitive'. A schedule to the Act lists some classes of information that are considered sensitive. Health information is included in this list.

3.22 In the event of a damaging identification or disclosure of information about an individual in a statistic, an inquiry by the National Statistician, the Statistics Commission, or a Court would establish whether all reasonable steps had been taken to prevent this. An inquiry can expect to find that some account had been taken of the degree of sensitivity of the information, and that the steps taken to protect the information reflects this sensitivity. Abortions information should be considered to be amongst the most sensitive of any information used for the publication of statistics, and therefore the test of 'reasonable steps' is a high one.

3.23 The practices for the use of abortions information is consistent with the data protection principles in the Data Protection Act (1998). The principles require that personal data are processed only where necessary. The personal abortions data are processed by DH only where necessary for the Department's statutory duties under the Abortions Regulations, and for certain research purposes within rigorous organisational and technical governance arrangements. All other uses of the abortion information, for example for the management of services in PCTs, are conducted using the non-disclosive statistics produced by Department of Health and ONS.

Application of the policy framework - protection for Abortion Statistics

The UN Fundamental Principles and the Code of Practice

3.24 Abortions statistics are aggregated from information in the notices of abortions made to the Chief Medical Officer. The content of the notice is specified in the Abortions Regulations, and the supply of the information is a statutory requirement. Section 5 of the Regulation prohibits the disclosure of information in the notices other than as set out in that section. A published statistic that was likely to disclose identifiable notice information should be considered a failure of the confidentiality guarantee in Principle 5 of the National Statistics Code of Practice. There would also be a failure against the 6th UN Fundamental Principle.

The Protocol for Data Access and Confidentiality

3.25 What does 'likely to disclose' mean in the context of abortion statistics? Applying the framework from the Protocol makes 'likelihood' a factor of who the intruders may be for these statistics, and what constitutes the disproportionate amount of time effort and expertise they may employ in an attempt to identify abortion notice information

Disproportionate time, effort and expertise:

- 3.26 Designers of abortion statistics should allow for the intruder in a relevant scenario to have access to data processing software and hardware equivalent in standard to that available to a Government Statistics Service practitioner in their usual job. The designer should allow for the intruder to have statistical and other relevant expertise equivalent in standard to those found in a Government Statistics Service Statistics Officer, e.g. to degree level. The designers should allow that an intruder of these skills and with appropriate software and hardware would be prepared to dedicate a number of hours of their time to the task of identifying an individual.

Other legal considerations

- 3.27 The Freedom of Information Act (FoI) (2000) gives the public a general right of access to information held by public authorities. The Act applies to any information in any permanent form. The public can request to have a copy of the information supplied to them, to have a summary of it provided, or to inspect the information in situ. In most circumstances, the information must be provided within 20 working days. Some classes of information may be exempt from these requirements, and if exempt, the Authority can withhold the information requested.
- 3.28 ONS has made guidance available to the Government Statistical Service. The primary consideration is to make a clear distinction between information that, if disclosed to a member of the public under the Act, would breach the confidentiality guarantee, and information that if disclosed would not. The exemptions in the Act must be used to preserve the confidentiality of statistical records. But information held by producers of official statistics that is not confidential should, in usual circumstances, be disclosed to the public on request.
- 3.29 For the production and dissemination of abortions statistics, the FoI Act will require all who hold abortions records and statistics to be particularly assiduous in correctly identifying and classifying the information. It must be clear to all officials in the department and other public authorities involved whether each item of information is likely to be exempt information, or likely to be available to the public on request. To be clear - all abortions information, records or statistics that are likely to allow for the identification of an individual abortion notice entry, or to reveal information about a notice entry not already in the public domain, must be withheld from the public using relevant exemptions. The status of the information or data should be clear to all practitioners, although sometimes it will not be obvious that the information does identify an individual.
- 3.30 Information scheduled for publication is exempt information. Statistical information provided to officials involved in the formulation and development of government policy may be exempt, but only until the decision as to government policy has been made. Statisticians and researchers may also engage in the free and frank exchange of views and provision of advice for the purposes of deliberation, and the information exchanged may be exempt information.
- 3.31 The Freedom of Information Act makes a distinction between factual information and statistical information. Statistical information is a summary, projection, or estimation, etc based on a sample of observed facts.
- 3.32 It may be useful to consider the distinction between factual information and statistics for the management of abortions information. The Regulations provide that the CMO can disclose abortions notice information (factual information) to a number of specified persons. Some persons to whom the notice information is disclosed shall use the information for the administration of abortions and related services. Where lawful, it may be appropriate and necessary for administrative purposes for the subjects in the

notice to be distinguished and treated separately or differently from the other notice subjects.

- 3.33 Others to whom the notice information is disclosed shall derive statistical information from it. Published statistics may be used for any purpose. Unpublished statistical information can be used for statistical and research purposes only. It is a breach of the Code of Practice to distinguish and treat separately or differently any one statistical unit in derived statistical information for a non-statistical purpose, or to allow for this to be likely in a published statistic. This puts limitations on the utility of statistics for the administration and monitoring of health services.

Specific ethical issues

- 3.34 It should be recognised when considering the balance between utility and disclosure risk that information in abortions notifications is particularly sensitive. Recent experience shows that such sensitive and restricted information may become the focus of high profile public debate.
- 3.35 The Department for Constitutional Affairs is conducting a review of statutory disclosure bars. The Freedom of Information Act allows for the Lord Chancellor to revoke or amend any absolute prohibition on disclosure. Many have been removed already, and a majority are to be amended to have instead a 'sunset clause' - a fixed number of years after which the information becomes available to the public. It is an illustration of the sensitivity of abortions data that the Department for Constitutional Affairs has in its review agreed that the bar for the abortions notices shall remain an absolute bar.

4. Risk Management

- 4.1 When managing the confidentiality issues associated with a data set a balance needs to be struck between the usefulness of the data and the disclosure risks. In order to achieve this balance a risk assessment of the data set must be implemented as well as gaining a good knowledge of the users and uses of the data. The next section provides details on the users and uses of the abortion data. A detailed risk assessment of the statistics then follows and leads to the formulation of principles for reducing the risk in the data set to an acceptable level.

Utility

- 4.2 There are a number of users and uses of the abortion statistics. The principal users are the Office for National Statistics (ONS), Department for Education and Skills (DfES), professional bodies such as the Royal College of Obstetricians and Gynaecologists (RCOG), professional groups such as researchers and service providers and Primary Care Trusts (PCTs) and Strategic Health Authorities (SHAs) in England and Local Health Boards (LHBs) in Wales, the Healthcare Commission and lobby groups. In general the data will be used to monitor and analyse trends in abortions. In particular the ONS use the abortion data and statistics to compile statistics on conceptions and to input into the Health Statistics Quarterly (HSQ), a publication covering the latest trends in the UK's health. PCTs, SHAs and LHBs will also use the statistics for planning and to commission services. In order to achieve this they will need statistics for the number and type of abortions that are performed in their area (such as gestation, ethnicity or repeat abortions) and information on funding.
- 4.3 Abortion statistics, as part of the wider conceptions statistics (births, abortions, and still births), are used for monitoring the effectiveness of government policy on reducing the rates of teenage pregnancy. The Social Exclusion Unit's 1999 Report on Teenage Pregnancy set national targets for the reduction of conceptions rates for under 18s and under 16s by 2010. The reduction of the under 18s rate by 50% by 2010 is a Public Service Agreement owned jointly by the Department of Health and Department for Education and Skills. The Priorities and Planning Framework 2003/6 translates these national targets into key priorities for health and social care, and this gives rise to the Local Delivery Plans for all PCTs/LHBs. Each PCT's/LHB's Local Delivery Plan has a trajectory for a reduction in under 18s conceptions rates.
- 4.4 A key test of practical utility, as required by the UN Principle, is whether the abortions (conceptions) statistics allow PCTs/LHBs to plan and commission services and to monitor their Local Delivery Plans for teenage pregnancy rates.

Risk assessment

- 4.5 In order to apply appropriate disclosure control methods and make the data safe for release it is essential to be explicit about what disclosure risks could be present in the data and need to be protected against. This section details this process in particular discussing who requires protection, what disclosure risks need to be protected against and what parts of the outputs pose a risk of disclosure.

Who are we protecting?

- 4.6 The National Statistics Protocol on Data Access and Confidentiality applies to all statistical units. Statistical units are defined as individuals, households or businesses. An abortion is an event related directly to an individual woman and the restricted details of the female involved in the abortion that are included in the notification form need to be protected. Protection of these details requires the protection of the identity of the household of which the female who had the abortion is a member. In the majority of cases this will follow directly from protecting the individual but additional measures may

need to be taken when different individuals from the same household are represented in the same cell, e.g. two sisters. In addition an abortion is an event that can be directly related to a practitioner. Therefore the identity of the practitioner as the person who carried out the procedure needs to be protected. Similarly the identity of the hospital/clinic as the place in which the termination occurred should also be protected. It follows that for abortion statistics a statistical unit represents an individual woman, a household, a practitioner or a hospital.

- 4.7 The hierarchical and cross-classified nature of the data and the links between the event (an abortion) and an individual woman, household, practitioner and hospital is outlined in Figure 1 below. One abortion is linked to one woman that is nested within household. More than one woman per household could be represented in the data, e.g. W1 and W2 both belong to H1. A woman could potentially have more than one abortion, e.g. A3 and A4 are linked to W3. Each abortion is linked to an operating practitioner who can perform more than one abortion, e.g. A1 and A2 are both linked to P1. Practitioners are nested within hospitals/clinics but can be linked to more than one, e.g. P3 works in both C2 and C3.

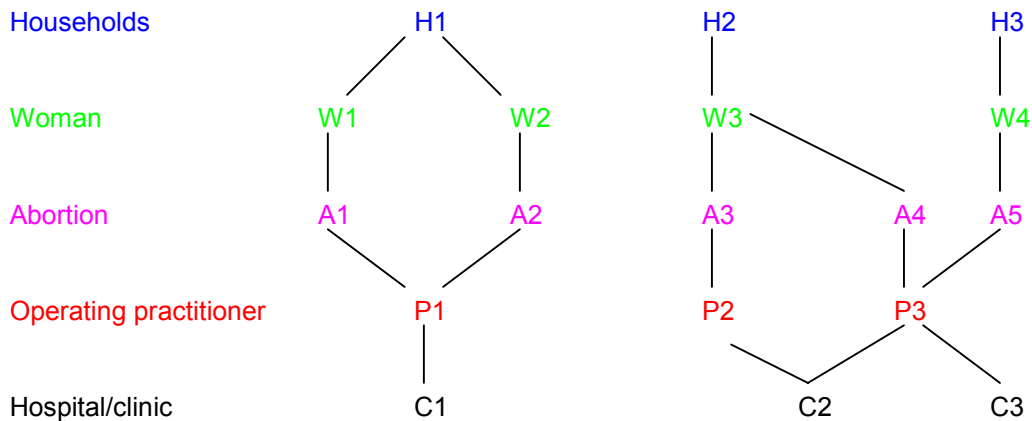


Figure 1: Data Structure

- 4.8 The number of different units within the data set and the complex relationships between them makes protecting the confidentiality different to standard cases where protection may only be required for the individual involved. The event or abortion is both the collection and output unit, but it is the woman, household, practitioner or hospital that require protection.

What disclosure risks need to be protected against?

- 4.9 In order to be explicit about the disclosure risks present in the data a range of disclosive situations are identified. Each situation or scenario is based on the behaviour of an intruder as introduced in Chapter 3 where the intruder selects or recognises by analysis the characteristics of a statistical unit and/or reveals information about the unit that is not already in the public domain. The Protocol considers three different types of disclosive situations: (i) self identification, (ii) identification of a statistical unit and (iii) disclosure of information not already in the public domain about a statistical unit. Type (ii) is termed identity disclosure and type (iii) is known as attribute disclosure. Different scenarios are identified under each type of disclosive situation.

Self identification

- 4.10 Here the intruder is represented in the data and selects or recognises by analysis the characteristics of him/herself. The Protocol only considers self-identification to be an

issue under a number of conditions. The Responsible Statistician will decide on release having considered all of these conditions, including:

‘where the information is personal, the statistical unit, following self-identification, could not reasonably claim that either the processing or the output could cause them substantial damage or substantial distress’

- 4.11 Due to the sensitive nature of abortion statistics the conclusion is drawn that self-identification of the woman having an abortion should be protected against since self-identification could lead to damage or distress.
- 4.12 **Scenario 1**
A woman knows some/all of the information about her that is in the table and can therefore find herself in a table. A similar scenario of self identification of a household also follows.
- 4.13 Due to the sensitive nature of abortions, the self identification of a practitioner performing such a procedure is also likely to cause damage or distress. Therefore self identification of a practitioner should be protected against. A similar scenario of self identification of a hospital also follows.
- 4.14 This concept of identification carries an element of uniqueness, the association of a name with a cell value. Thus there is a difference between being able to say that someone belongs to a population in a cell with a value of say, 162, (which must be non-disclosive), and being able to say that a particular named person is the individual in a cell with a value of 1.
- 4.15 One could argue that attribute disclosure as well as identification has occurred for this scenario since the table has disclosed that the individual is unique. Uniqueness in a 100% sample output allows individuals both to identify themselves, and to deduce that no-one else at all shares their combination of characteristics within the population of the table. This may be of concern to them, particularly if these unique characteristics are visible and/or sensitive.

Identification of a statistical unit

- 4.16 The outputs should not allow an intruder to identify a statistical unit, even if all/some of the details about that unit in the table are already known. Here protection is required to prevent an intruder selecting or recognising by analysis the characteristics of a particular individual statistical unit. Again the concept of identification carries an element of uniqueness, the association of a name with a cell value.
- 4.17 **Scenario 2**
Typically the identification is made by someone who already knows all/some of the details about the termination. The table allows identification and possibly disclosure of the fact that the abortion is unique, e.g. family and friends of a female who has had an abortion, know her well and know enough information about her and the abortion that they can find her in the table. The friend or family here may be an undesirable intruder, e.g. an estranged partner. Note again that the unit being identified could also be a household, practitioner or hospital.

Disclosure of information (attribute disclosure)

- 4.18 This disclosive situation involves the intruder revealing information about the statistical unit that is not already in the public domain.
- 4.19 **Scenario 3**

Someone who already knows a female who has had an abortion knows some of the information about her that is in the table, can find her in the table and could disclose further information, e.g. the father of a female knows her age, where she lives and that she had an abortion. He identifies her in the table and discovers the gestation period or the statutory grounds for the abortion. Again the intruder may be undesirable. This scenario can also be applied to households.

4.20 Scenario 4

An intruder knows some information about a practitioner that leads to an identification from the table and a disclosure of further information, e.g. the intruder discovers that the practitioner performs terminations on certain grounds.

4.21 Scenario 5

An intruder knows some information about a specific hospital, can find it in a table and could disclose further information, e.g. the intruder discovers that the hospital carries out a specific type of procedure.

4.22 Scenario 6

An intruder who knows some information about a female who has had an abortion tries to find her in the table. They can not find her in the table (some of the information that they know is not true) and thereby disclose something about her, e.g. a woman told her husband that she had a termination on certain grounds, but the table reveals that none of these took place. This scenario would also hold for households, practitioners and hospitals.

4.23 Scenario 7

Identification of an individual or practitioner and further disclosure caused by using the statistics in conjunction with other information. For example an intruder uses the information in the table to find the name (and/or other information) of a person they didn't already know, e.g. a local journalist discovers from the table that there is only a small number of people in a particular cell. The small number in the cell doesn't tell the journalist who the individuals are but it may prompt him to follow up private sources of information to locate the individuals and thereby disclose further information. The likelihood of this scenario is increased for more sensitive cells, by the presence of small counts and for smaller geographical regions.

4.24 Note, a scenario that was considered but discounted involved the situation where disclosure occurs due to a 100% count in a cell e.g. all women of a particular age in a particular region/neighbourhood had an abortion. The scenario was discounted since the likelihood of it occurring is remote and a woman can potentially have multiple abortions in the time period covered by the data introducing sufficient ambiguity in the count considered. This likelihood is linked to the geographical level at which the statistics are published, therefore care should be taken if abortion statistics are published at lower levels.

What parts of the outputs pose a risk of disclosure?

4.25 The disclosure scenarios outlined above can now be used to identify the parts of the outputs that pose a risk of disclosure, usually associated with low counts or data for small areas. Where a cell within a table could lead to a disclosure risk, i.e. identification or further disclosure of information, that cell will be described as an unsafe cell. Once identified, the risk of disclosure will need to be reduced by disguising all those cells considered to be unsafe.

4.26 Scenario 1 and 2

For frequency table outputs from "whole population" data sources, the identification in scenario 1 and 2 is immediate in any cells with a count of 1, representing one abortion, one event. A cell of size 1 identifies a unique termination (apart from data quality). No

effort is required beyond an interest in understanding the table. Therefore cells with a count of 1 are considered unsafe. This might be extended to cells with a value of 2 representing two events, where one of the individuals contributing to one of the events in the cell may identify the other individual contributing to the events in the cell. For example, two women may have met in the hospital and had some form of discussion about their termination. Either patient seeing a 2 in the table would be able to see themselves and the other patient uniquely. This is much less likely than identification following publication of a 1 and the reasoning is not normally extended beyond cells of size 2. Therefore scenario 1 and 2 result in cells of size 1 and 2 being classed as unsafe.

- 4.27 Consider the table of statistics shown below, where the number of abortions of type 1 and 2 is broken down by age bands. The woman who is under 12 and had a type 1 abortion will immediately be able to recognise herself in the table.

Outcome	Age				Total
	< 12	12-15	16-19	> 19	
Type 1	1	15	7	3	26
Type 2	0	7	18	19	44
Total	1	22	25	22	70

- 4.28 In some circumstances self identification or identification could occur from cells with counts greater than 2. It is possible (as shown in Figure 1) that one woman can be associated with more than one event, e.g. a woman can potentially have more than one abortion during the period covered by the statistics. Cells where all events in that cell referred to the same individual should not be published as this could lead to scenario 1 and 2. To protect against this at least 3 different individuals should be represented in each cell. When there are 2 individuals this prevents one individual identifying the second. Therefore any cell in which the count of events is associated with either 1 or 2 women is considered unsafe.
- 4.29 Similarly in order to protect the identification of households any cell in which the count of events are associated with either 1 or 2 households is considered unsafe. For example suppose the table above represents statistics for a small area and it just so happens that two sisters (over 19 years old) belonging to the same household have abortions (all of type 1) in the same year and one sister has two abortions. Then the 3 in the cell for >19 and type 1 will be dislosive in this situation.
- 4.30 Protection for this type of disclosure could be implemented by referring back to the individual micro-data and checking the number of different unique individuals or households contained in each cell in all tables. This process would be extremely resource intensive. Relating back to the principle in the Code of Practice that users' needs are to be met 'within available resource' a simpler approach can be to consider all cells of size $<n$ terminations as unsafe. A judgement is required to determine n balancing the risk of identification and the usefulness of the data. The likelihood of a cell containing 5 or more events that relate to only 1 or 2 women or 1 or 2 households is considered to be small. Therefore in order to protect against scenario 1 and 2 for women/households cells of size <5 are considered unsafe.
- 4.31 Scenario 1 and 2 can also involve the identification of a practitioner as well as an individual who has had a termination. As for individuals and as shown in Figure 1 it is possible that the count of events represented in any specific cell can refer to work done by a single practitioner or at a single hospital. So any cell in which the count of events is associated with either 1 or 2 practitioners/hospitals is considered unsafe.

- 4.32 In the example table suppose that all abortions of type 2 for 12-15 year olds were carried out by the same practitioner. From the 7 in this cell the practitioner could make a self identification or be identified by others.
- 4.33 One hospital/practitioner can perform a large number of procedures and so a simple threshold as applied to protect individuals is not appropriate. For the majority of cells the likelihood of the count of events being associated with 1 or 2 practitioners/hospitals is low. However for some cases, e.g. certain medical procedures or abortions performed over 24 weeks, only a small number of practitioners and hospitals might perform these procedures. Therefore the likelihood of these cells being unsafe is higher. In addition these cells will have lower counts generally, are the more sensitive procedures and are more likely to attract the interests of intruders. Therefore for the types of abortions listed above unsafe cells will be identified by referring back to the individual micro-data and checking the number of different practitioners/hospitals contributing to the count in each cell.
- 4.34 **Scenario 3, 4 and 5**
These scenarios involve the intruder revealing information not already in the public domain. In a table, identification will occur from a 1 in a marginal total that will then reveal further information within the row or column. A marginal total is a total within a table, i.e. a row or column total. For example, the table shows a 1 in the <12 column total. If an intruder knows a girl who is under 12 and has had an abortion the table then immediately reveals that this was a type 1 abortion.
- 4.35 Using the argument as above, in order to protect against these scenarios any 1's or 2's from marginal totals need to be disguised. In addition marginal totals should represent more than 2 individuals, households and more than 2 practitioners/hospitals.
- 4.36 Disclosure of information could also occur under this scenario from zeros in a table. Suppose an intruder knows the place of residence and the age of a woman having an abortion. The table reveals that all abortions performed in that area and for that age group were of a particular type since all the cells for other types are zeros. This leads to the conclusion that all non-structural zeros within the table are unsafe. A zero is structural when a count in a cell is impossible, e.g. some procedures are not possible for late gestation. In the example the table reveals that all women having type 2 abortions are 12 years or older.
- 4.37 **Scenario 6**
Scenario 6 also leads to the conclusion that all non-structural zeros within the table are unsafe.
- 4.38 **Scenario 7**
This scenario requires some effort to find the name or more details about the individual, household, practitioner or hospital. In a large population (e.g. country or GOR) the effort and expertise required may be deemed to be disproportionate, as outlined in 3.1. As the base population decreases to smaller geographies or sub-populations it becomes easier to find units. A judgement is required to determine what is an unsafe cell for this scenario. At a minimum all cells containing a count of 1 or 2 for smaller geographies are considered unsafe.
- 4.39 Abortion data are very sensitive, attract a lot of attention and therefore the impact of any identification or disclosure from these statistics is considered to be very high. All the data are considered sensitive and so for this scenario an appropriate level is to consider all cell counts <5 as unsafe.
- 4.40 The effort and expertise required by an intruder to make an identification will decrease for smaller geographies and so for tables published for PCO/LHB and SHAWelsh regions any cells of size <10 will be considered unsafe. So for the majority of cells

reported for England, Government Office Regions and Wales the effort and expertise required by an intruder to make an identification is considered disproportionate for counts of size 5 and above whereas for smaller geographies any cells of size <10 will be considered unsafe.

- 4.41 For variables considered highly sensitive the interest in the numbers and the value of an identification or disclosure will be increased. Therefore the effort that an intruder will go to in order to disclose information or make an identification is increased. For this reason for tables containing those particular variables any cells of size <10 will be considered unsafe at whatever geography.

The variables that are considered highly sensitive are:

- Young ages (<15)
- Late gestation (over 24 weeks)
- Procedure by gestation
- Medical conditions

- 4.42 Where a table contains any of the above variables then all the cells within that table are considered unsafe if the count is <10.

- 4.43 Introducing disclosure rules for the sensitivity of variables or the value of the information to a journalist does introduce an element of subjectivity that may cause difficulties for a data supplier. In the majority of cases a minimum and standard protection level will suffice but since abortions are considered to be one of the most sensitive of health statistics then the additional protection is appropriate, see Chapter 3. These decisions on sensitivity have been made in discussion with those who have a detailed understanding of the statistics and experience of the high profile public debate and interest in the figures.

- 4.44 Each scenario described above has been used to identify the parts of the abortion statistics that pose a disclosure risk. All of these conclusions can be combined to define unsafe cells within the outputs.

- 4.45 **Conclusion 1: Within a release of abortion statistics unsafe cells are defined as being counts of abortions that are:**

- **zero unless no other value is logically possible**
- **less than 5 for Government Office Region in England, the country of Wales or any larger geography**
- **less than 10 for any geography below the Government Office Region in England or the country of Wales**
- **less than 10 for highly sensitive variables**
- **associated with either 1 or 2 practitioners**
- **associated with either 1 or 2 hospitals**

The variables that are considered highly sensitive are:

- **Young ages (<15)**
- **Late gestation (over 24 weeks)**
- **Procedure by gestation**
- **Medical conditions**

Other data sources

- 4.46 Chapter 4 identified the parts of the abortion data that pose a risk of disclosure. The Protocol on Data Access and Confidentiality states that the confidentiality guarantee will be met by:

'taking account of information likely to be available to third parties'

- 4.47 At one level this has been covered by scenario 7 but another stage involved in risk assessment of the abortion data is to consider other data sources that are available that might contribute to disclosure risk. This fits into three broad categories covered below.

Information published about abortions

- 4.48 Linking together different tables in the release should not lead to disclosure. Where tables provide data in terms of rates or percentages the numbers themselves may not be disclosive. However, when combined with other tables it may be possible to recover the original counts, which may be disclosive. If the rate or percentage is based on an unsafe cell and it is possible (by linking with other tables in the release) to recover the original count then the cell with the rate or percentage is itself unsafe. Some protection can be provided by rounding rates or percentages. However, care still needs to be taken to avoid disclosure. Protection will be provided if the base from which the rate or percentage is calculated is sufficiently large since the implied count could be a range of values, however, this range must be large enough to satisfy disclosure rules and thresholds.

- 4.49 **Conclusion 2: Simple calculations such as rates or percentages do not necessarily make an unsafe cell safe and should not be used to protect data unless it can be demonstrated that one cannot work back to the original count. In order to keep statistical disclosure control rules clear and consistent for abortion statistics rates or percentages should only be calculated from safe cells.**

Information that could be published about abortions

- 4.50 Care should also be taken when publishing abortion statistics in any other format which could lead to disclosure. Of concern would be a request to publish counts by a geography other than PCO/LHB/SHA/regions. This data may not be disclosive but could be differenced with the published data. Depending on the method used to protect the data the resulting differenced table could be disclosive. Care should also be taken when publishing data by non-standard categories e.g. non-standard age ranges (for example if the standard age band was 18-45 then the number of terminations in the age range 19-45 could permit us to difference the number of terminations for 18 year olds).

Other freely available datasets which could lead to disclosure.

- 4.51 The abortion rates are published by PCO/LHB and age per 1000 in the population. If population estimates for females in the same age bands are available then these rates could be used to calculate potentially disclosive counts. As above a rate or percentage calculated from an unsafe cell is also unsafe if the original count can be recovered.

Other datasets which include abortion data

- 4.52 The ONS use the abortion data to compile statistics on conceptions. The publication of these statistics must not allow any unsafe cells from the abortion statistics to be revealed. This section examines this likelihood.
- 4.53 In order to make any disclosive inferences about abortions from counts or rates of conceptions at National, SHA/region, County/Unitary Authority (UA) or LA level an intruder will need to have access to counts of maternities (live + still) aggregated by the same age bands and for the same geographical level as for the conceptions data. Due to the time lag between conceptions data and births/abortions data, the unknown number of multiple births and the differences in the age of the mother at birth and at

conception, the count obtained by subtracting maternities from conceptions is highly unlikely to reflect the true number of abortions.

- 4.54 At ward level no data are available on stillbirths and live births are only reported for all ages. An intruder could difference the total number of conceptions and the total number of live births to arrive at an estimate of the total number of abortions + stillbirths for the ward. Protection against disclosure of the number of abortions will be provided by the unknown number of stillbirths, multiple births and the time lag between conception and birth/abortion. The number of abortions for under 18s cannot be derived from the conception statistics since data on stillbirths and live births is not available for this age band.
- 4.55 Outcome rates and percentages of conceptions leading to abortions are also published. Although the figures are rounded in many cases the original count of abortions can be estimated. Although protection against disclosure of the number of abortions will be provided by the unknown number of stillbirths, multiple births, differences in age at conception and birth, and the time lag between conception and birth/termination the effort and expertise required by the intruder to obtain an estimate is not great. Therefore more protection is required. If the count of abortions estimated from the outcome rate or percentage conceptions leading to abortions is considered unsafe then in order to be consistent with the abortion statistics such cells will need to be disguised in the conceptions publication.
- 4.56 Abortion statistics are published by SHA/region and PCO/LHB geographies whereas conception statistics are published by SHA/region and LAs. PCO/LHB are non-coterminous with LA. Some further protection of the LA outcome rate or percentage conceptions leading to abortions figures may be required to cover the case of disclosure by differencing with PCO figures. More technical details of this risk assessment are contained with Appendix 2.
- 4.57 The conclusion is drawn that conception statistics displayed as counts and rates cannot be used to disclose information on abortions. Therefore counts and rates should continue to be published using current disclosure control methods. Where figures are published for percentage conceptions leading to abortions and rates of conceptions leading to abortions protection will be required where the derived number of terminations is considered unsafe (e.g. less than 5 or less than 10). Some further protection of the LA outcome rate and percentage conceptions leading to abortions may be required to cover the case of disclosure by differencing.

Data Quality and Coverage

- 4.58 In order to assess the disclosure risk of a data set an assessment of the coverage and quality of the data must be considered. If the quality of the data is known to be poor then some ambiguity will be introduced into the counts offering some protection against disclosure.
- 4.59 The abortion statistics cover all legal abortions that are carried out by a registered medical practitioner in England and Wales. The Abortions Act requires that any termination must be notified within fourteen days to the Chief Medical Officer of the Department of Health (DH) or to the Chief Medical Officer of the National Assembly of Wales. Since the process of notifications is set out within the Act then the quality of the information provided in the notification is considered high. In addition the DH carry out quality checks on the data and make enquiries in order to fill in any missing data items. Since the quality of the data is considered to be high then no protection of the data can be assumed from a quality point of view. There can be no assumption that there is variability which would protect against identification of data in unsafe cells.
- 4.60 A general rule for protecting confidentiality is that the more recent the data, the stricter the application of disclosure control must be. In some circumstances, allowances can

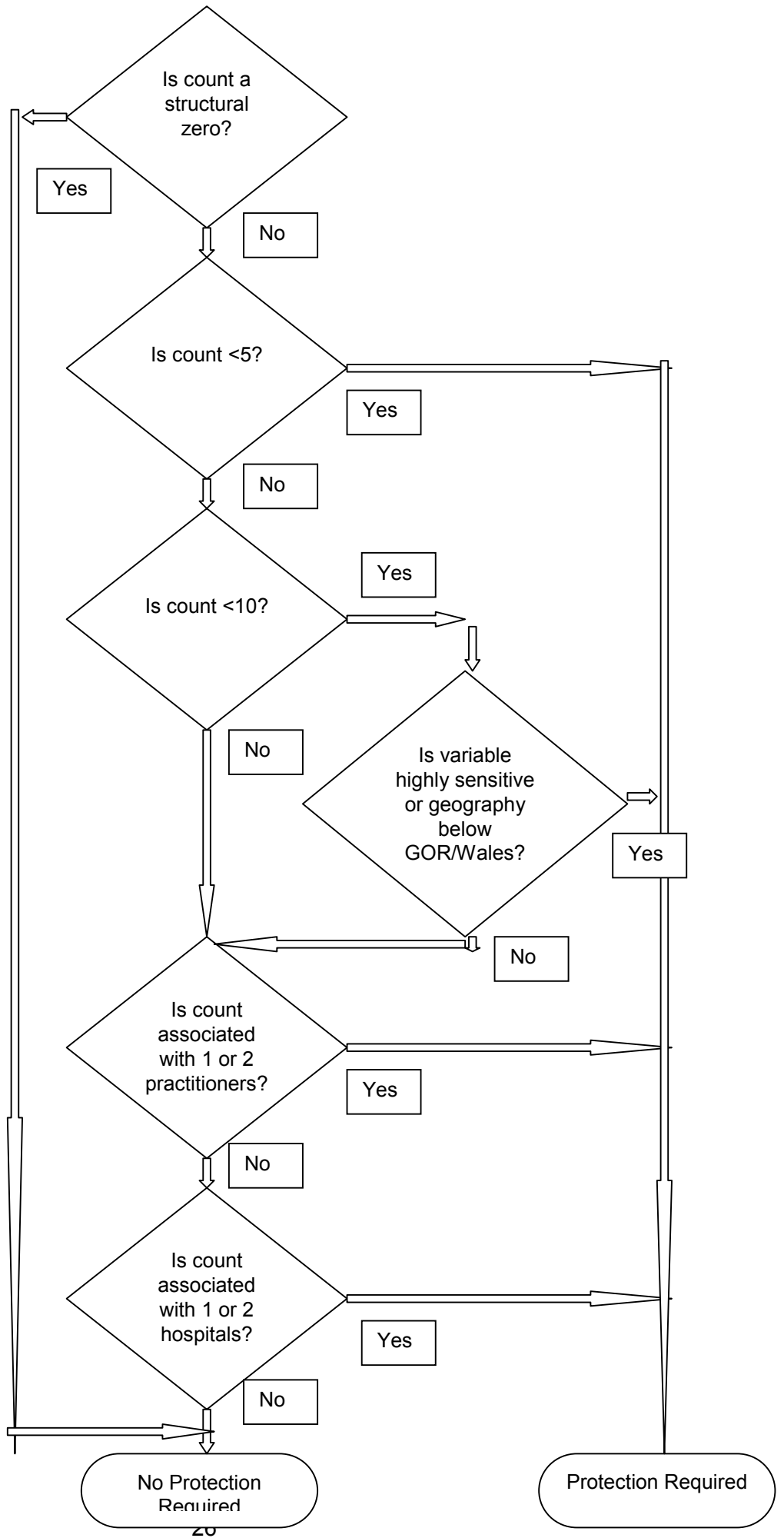
be made for any reduction in the likelihood of identification as the data get older. These abortion statistics relate to 2003 and 2004 and so no such allowances are made. Another aspect of data quality to be considered when protecting the confidentiality of statistics is the impact of the disclosure control methods on the overall quality of the statistics released. The different methods will all summarise, modify or perturb the data in some way in order to disguise the unsafe cells. The impact on data quality of the different disclosure control methods is covered in Chapter 5.

Flow diagram

A flow diagram below shows graphically the conclusions made in this Chapter on risk assessment and how they should be applied to protect the confidentiality of the abortion statistics.

Figure 2: Risk Management Principles – Flow Diagram

Highly sensitive variables:
 Young ages (<15)
 Late gestation (over 24 weeks)
 Procedure by gestation
 Medical conditions



5. Disclosure Control Methods

- 5.1 The previous Chapter of the report provided a detailed assessment of the risks associated with the abortion statistics as well as information on how the data are used. As outlined in Chapter 2 only non-disclosive data that are sufficiently abstract from the information provided in the notification can be published. As introduced in Chapter 3 the Code of Practice states that no statistics will be produced that are 'likely to identify' a statistical unit. In order to ensure that these specifications are met this Chapter describes the disclosure control methods that can be used to disguise cells considered to be unsafe while balancing the risks with data utility. The risk within the data is not entirely eliminated but is reduced to an acceptable level. A description of different disclosure control methods that could be used is provided in Appendix 3 and here the advantages and disadvantages of implementing each for the abortion statistics are outlined. This includes the protection and the amount of damage done to the data.

Preliminary Protection methods

- 5.2 Preliminary protection methods are used to design tables that minimise the number of unsafe cells or modify the data pre-tabulation.

Table Redesign

- 5.3 For a particular variable unsafe cells can be disguised by collapsing categories, e.g. producing a table with counts displayed in age bands rather than single year of age. The advantages for disguising unsafe cells within the abortion data using table design is that the original counts in the data are not damaged. However, the detail displayed within the table will be reduced. The method is easy to implement but does require a good knowledge of the data and an awareness of the needs of users in order to combine categories whilst maximising the utility in the data.

Geographic thresholds

- 5.4 If tables contain too many unsafe cells, then one solution is to increase the frequency count for each cell by aggregating to a higher level of geography. This method is very straightforward to implement. Apart from the loss of detail, the data need not be damaged: the published frequencies maintain their correct values. The risk of identification is reduced since the individual frequencies will be larger and the population at risk for the intruder to search is also increased, as is the geographical spread.
- 5.5 There may however, be sound policy or other practical reasons for wanting to output at a particular geography, which would then call for alternative disclosure measures to be selected.

Place of Residence

- 5.6 As described in Chapter 4 cells in which the count of events is associated with either 1 or 2 practitioners/hospitals are considered unsafe. One method to reduce the number of such unsafe cells would be to construct the statistics by area of residence of the individual woman rather than by place of termination. The advantage of this method is that it will significantly reduce the risk from this type of disclosure but some cells still may be unsafe. A disadvantage of the method is that data by place of termination will not be available.

Combining years

- 5.7 Another preliminary protection method that could be used to protect unsafe cells would be to aggregate tables across a number of years, e.g. 3, thus reducing the number of

unsafe cells. Data could either be published every 3 years, .e.g. publish 2003-2005 in 2005 then 2006-2008 in 2008, etc or as a rolling total every 3 years, e.g. publish 2003-2005 in 2005 then 2004-2006 in 2006, etc. A technical difference between the disclosure risks associated with these two methods is provided in the appendix.

- 5.8 This method could be used to provide more detailed data for particular tables but needs to be considered in light of users' needs. In particular PCTs, SHAs and LHBs require data on an annual basis for planning services. The number of years aggregated will be determined by the number of unsafe cells which are defined in the same way as for annual data.

Record Swapping

- 5.9 A pre-tabulation disclosure control method that could be considered is record swapping. This method involves swapping pairs of records that are partially matched (e.g. individuals that have the same age) which alters the geographic locations attached to the records, but leaves all other aspects unchanged. The disadvantage of implementing record swapping for the abortion statistics is that a high level of swapping would be required in order to disguise all unsafe cells. The distribution of the statistics within tabulations produced from the record-swapped data would be distorted and a user would not be aware of the level or type of distortions. In particular counts within certain geographical areas would be perturbed. This would be highly undesirable for the users of abortion statistics.

Further protection methods

- 5.10 The statistical disclosure control methods outlined in the previous section are essentially preliminary methods of disclosure control. If unsafe cells still exist in the output then further protection methods should be considered in order to disguise them. These further protection methods essentially involve a choice between suppression and rounding.

Suppression

- 5.11 A method of protecting unsafe cells in tables is cell suppression. This means that unsafe cells are not published, they are suppressed and replaced by a special character, such as '.' or 'X', to indicate a suppressed value. Such suppressions are called primary suppressions. To make sure that the primary suppressions cannot be derived by subtractions from published marginal totals, additional cells are selected for secondary suppression. The selection of secondary suppressions can be done either by hand or by software. More details on suppression can be found in Appendix 3.
- 5.12 Where the number of unsafe cells in a table is very low, e.g. 1 or 2, suppression will be a relatively easy method to implement and will not result in high information loss. A disadvantage of this method is that most of the information about the original values in the suppressed cells is removed and due to secondary suppressions some counts that are safe will also be removed. If the number of primary suppressions is not low then the information loss can be high and the ideal choice of secondary suppressions is not a trivial task. Another disadvantage of suppression is that in order to disguise unsafe zeros these cells will need to be suppressed which could again result in high information loss.
- 5.13 An introduced in Chapter 4 an additional disclosure risk could occur from disclosure by differencing if abortion statistics are released on non-standard geographies or categories. The disadvantage of suppression is that this method does not offer a solution to disclosure by differencing. This would mean that without a detailed analysis of disclosure by differencing the statistics could not be published on any other geographies other than SHA/region/PCO/LHB or with other non-standard variable

categories, e.g. age ranges. Therefore a careful audit process would need to be implemented for any tables released on an ad-hoc basis, this could be time consuming and therefore resource intensive.

- 5.14 An advantage of suppression is that the method will disguise all unsafe cells as set out in Chapter 4, including cells in which the count of events is associated with either 1 or 2 practitioners/hospitals.

Rounding

- 5.15 Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell while adding a small but acceptable amount of distortion to the data. Two alternative rounding methods are outlined in Appendix 3: random rounding and controlled rounding. In each case there is a choice of the base for rounding - common choices are 3 and 5. All rounded values (other than zeros) are then integer multiples of the chosen base.
- 5.16 The advantage of using rounding is that if the number of unsafe cells is large then the table can be protected while still providing counts for all cells. In general the information loss will be lower for rounding than suppression. Rounding offers protection from disclosure by differencing since the difference between two rounded tables will also be rounded. This means that protection using rounding offers more flexibility in outputs compared to suppression. Rounding will also protect zeros without removing them since for example within a table rounded to base 5 a zero could represent any count between 0 and 4.
- 5.17 Random rounding is relatively easy to implement however in some instances the protection can be unpicked. In order to assure adequate protection, the resulting rounded table needs to be audited. After applying random rounding there may be inconsistencies in data within tables (i.e. rows or columns do not add up) and between tables (i.e. the same cell is rounded to a different number in different tables). Controlled rounding preserves additivity within tables and works for hierarchical data and for linked tables, however the method needs to be implemented in a statistical disclosure control software package.
- 5.18 A disadvantage of rounding for protecting abortion statistics is that there are difficulties in disguising cells in which the count of events is associated with either 1 or 2 practitioners/hospitals. If a cell had an original count of 17 cases all carried out by one practitioner, then rounding this to 15 means that the count still relates to only one practitioner, the unsafe cell is not disguised. A further disadvantage of rounding is that some users require exact counts for these statistics and rounded values would not be appropriate.
- 5.19 A further disadvantage of using rounding to protect the abortion statistics is that different bases would be required for different tables where the <5 and <10 rules are applied. This could result in inconsistencies between tables in the release and could lead to a disclosure risk.
- 5.20 A post tabular method similar to rounding is called Barnadisation and is a form of random record swapping for frequency tables where internal cells of every table are adjusted by +1, 0 or -1 according to probabilities. Typically the majority of cells are not modified. As in most post-tabular adjustments it leaves inconsistent totals between tables. In addition, as in random record swapping, it leaves high risk in the small cells, i.e. the probability that a 1 is a true 1 is quite high.

Alternative methods for Presenting Data

- 5.21 Alternative methods for presenting data can be considered as an approach for providing users access to information without disclosing the underlying data. This could include presenting data graphically or providing commentaries or analytical outputs. However, care needs to be exercised to ensure that the outputs are safe, for example scatter plots do not allow the identification of outlying data points.
- 5.22 As discussed in Chapter 4 data can be displayed as rates or percentages and some protection provided by rounding the final figure. Protection will also be provided if the base from which the rate or percentage is calculated is sufficiently large. However, one must ensure that the implied count satisfies disclosure rules and setting a minimum value for denominators may lead to safe cells being hidden. In order to keep disclosure rules clear and consistent it is easier to only calculate percentages or rates from safe cells.

Statistical Disclosure Control Summary

- 5.23 Chapter 4 outlined the cells within the abortion statistics that must be protected. Although redesign will reduce the detail within a table the method will protect many cells, it is easy to implement and does not damage the original counts in the data
- 5.24 **Conclusion 3: In order to ensure that unsafe cells in the abortion statistics are disguised table design should be used as a preliminary protection method. Redesign should be implemented taking into account the information required by the main users of the data.**
- 5.25 **Conclusion 4: Statistics should only be constructed using area of residence rather than place of termination thus reducing the risk of disclosure from counts of events that are associated with 1 or 2 practitioners/hospitals.**
- 5.26 The proposed annual bulletin release provides abortion data by place of residence for England and Wales and some information at the Strategic Health Authority (SHA) and Primary Care Organisation (PCO) level in England and regions and Local Health Boards (LHB) for Wales. The number of unsafe cells after table redesign should be minimised and so the information loss from suppression not significant.
- 5.27 **Conclusion 5: If unsafe cells exist in tables after redesign these should be removed using suppression methods (primary and secondary).**
- 5.28 Although suppression will limit the flexibility of outputs in particular ad-hoc releases it will disguise all unsafe cells including cells in which the count of events is associated with either 1 or 2 practitioners/hospitals. Since the number of suppressions is very low the method will be relatively easy to implement and offer a clear and consistent solution to the problem.
- 5.29 **Conclusion 6: In order to avoid resource intensive analysis of disclosure by differencing the abortions data should not in general be published on geographies that are non-coterminous with SHA/region/PCO/LHB or for non-standard variable categories.**
- 5.30 **Conclusion 7: In order to release more detail some data should be published aggregated over a number of years. To keep the methods for disclosure control clear, consistent and easy to implement the conclusion is made that rolling aggregates are not produced but years are aggregated independently.**
- 5.31 Current practice for disclosure protection for abortion statistics was described in Chapter 2. The implementation of the conclusions above will not significantly change the proposed release. Some additional table redesign or suppression will be required to protect counts associated with either 1 or 2 practitioners/hospitals, non-structural zeros

and percentages and rates based on unsafe cells. More detailed information published last year for the 2003 data can now be released on abortions where the threshold for disclosure has been reduced from 10 to 5 and where data aggregated for a number of years will be supplied.

6. Findings

Implementation

- 6.1 The aim of this report was to provide guidelines for interpreting the Code of Practice and associated protocols in relation to abortion statistics. Using these principles the parts of the abortion statistics that pose a risk of disclosure (termed unsafe cells) have been identified. Once identified conclusions have been drawn on how to disguise these unsafe cells and reduce the risk of disclosure.
- 6.2 These conclusions are made for the release of abortion statistics and are particular to the risks, utility and design of the data.
- 6.3 The guidance is written for those within the health community who are involved in the release and publication of abortion statistics. The guidelines will be mandatory for the Office for National Statistics. The intention is that they will be adopted by the Department of Health and The Health and Social Care Information Centre. It is hoped that others within the health community will also adopt the guidelines.
- 6.4 **Conclusion 8: The guidelines should be implemented for all published outputs of abortion data e.g. in cases where the data published can be used to make direct inferences about abortions. The guidelines should be implemented as soon as possible. The DH and ONS should work together in order to implement these guidelines for the annual bulletin release for 2003 and 2004.**
- 6.5 Note that keeping statistical disclosure control rules and methods consistent over time has considerable advantages. The guidance provided is based on current best practice. An operational review of disclosure control standards and methods, consistent with the National Statistics review process, should be carried out in association with ONS after a period no longer than 5 years.

Can data required by users be released?

- 6.6 As described in Chapter 4 the principal users of the abortion statistics are PCTs, SHAs and LHBs that need the data for planning, to commission services and to monitor their Local Delivery Plans for teenage pregnancy rates. A key test of practical utility of the conclusions above is whether the outputs will allow the PCTs, SHAs and LHBs to fulfil this role.
- 6.7 Detailed information on abortions by age, marital status, purchaser, statutory grounds, procedure, ethnicity, parity, complications, gestation period, medical conditions and non-residents will be published at the national level. Information on age, purchaser and gestation will also be provided for PCOs/LHBs and SHAs annually, providing the data needed for planning and to commission services. In a very small number of cases some information may need to be suppressed, however where this and other detailed information is considered unsafe at this geographical level then data will be supplied aggregated over a number of years.

Publication of Conceptions

- 6.8 As outlined in Chapter 4 the conclusions for this review will not effect the statistical disclosure control requirements for the conceptions statistics displayed as count data. For conceptions leading to abortions and rates of conceptions leading to abortions which are published at National, SHA/region, County/UA and LA level protection will be required where the derived number of terminations is considered unsafe (e.g. <5 or <10). Some further protection of the LA rate and percentage conceptions leading to abortions may be required to cover the case of disclosure by differencing. This may

lead to a small number of suppressions of these figures particularly in the smallest authorities.

- 6.9 This guidance has considered the disclosure control requirements for abortion statistics and for conception statistics where the data can be used to make direct inferences about abortions. It is not within scope to consider the disclosure control requirements for all conception statistics just to isolate any particular restrictions that result from the conclusions outlined above.

Release information

- 6.10 The parameters in any method used to disguise unsafe cells in a statistical release should in general be kept confidential by the data-supplying agency. Detailed knowledge about the disclosure control method applied can sometimes help intruders to undo the techniques used on some datasets. It is recommended only to state which technique(s) have been applied. Such a statement is important for two reasons. The first concerns the public perception of disclosure: declaring the use of disclosure control provides public confirmation that a dataset has been assessed for disclosure risk, and that methods of protection have been applied. Secondly, for quality purposes: the Protocol states that the users of a dataset will be provided with an indication of the nature and extent of any damage after disclosure control methods have been applied. Users will want to know how to interpret the figures.
- 6.11 **Conclusion 9: Users should be made aware of what constitutes an unsafe cell within the abortion statistics. The user should also be told that the method used to protect the table is predominantly table redesign used to minimise the number of unsafe cells that require suppression. The impact on the quality of the data will be that in some cases less detail will be displayed and suppressions will mean that some information is removed from the table.**

Appendix

Appendix 1 – List of Tables

A.1.1 This appendix contains a list of the tables proposed for release in the abortion statistics annual bulletin.

- Table 1. Legal abortions: resident status and purchaser, 1968 to 2003.
- Table 2. Legal abortions: age by (i) purchaser, (ii) statutory grounds, (iii) gestation weeks, (iv) procedure, (v) marital status, (vi) ethnicity, (vii) parity, (viii) previous miscarriages, (ix) previous abortions, (x) chlamydia screening, 2003
- Table 3. Legal abortions: by (i) purchaser, (ii) statutory grounds, (iii) gestation weeks, (iv) procedure, (v) marital status, (vi) ethnicity, (vii) parity, (viii) previous miscarriages, (ix) previous abortions, 1993 to 2003.
- Table 4. Legal abortions: by age, 2003.
- Table 5. Legal abortions: gestation weeks by purchaser, 2003
- Table 6. Legal abortions: duration of stay (nights) by purchaser and gestation weeks, 2003
- Table 7. Legal abortions: gestation weeks by age and purchaser, 2003
- Table 8. Legal abortions: procedure by gestation weeks, 2003
- Table 9. Legal abortions: complication rates by procedure and gestation weeks, 2003
- Table 10. Legal abortions: principal medical condition for abortions performed under ground E, 2003
- Table 11. Legal abortions: by strategic health authority and age, 2003
- Table 11a. Legal abortions: by primary care organisation and age, 2003
- Table 12. Legal abortions: purchaser and gestation by strategic health authority, 2003.
- Table 12a. Legal abortions: purchaser and gestation by primary care organisation, 2003.
- Table 13. Legal abortions for non-residents: by (i) country of residence, (ii) age (iii) statutory grounds, (iv) gestation weeks, 2003
- Table 13a. Legal abortions: country of residence by age and gestation, 2003

Appendix 2 – Conceptions Statistics

- A.2.1 This appendix examines the likelihood that statistics on conceptions will allow any unsafe cells from the abortions statistics to be revealed. This will determine whether the conclusions of this review will affect the statistical disclosure control requirements for conception statistics.
- A.2.2 Conception statistics are compiled by bringing together records of birth registrations, and of abortions performed under the 1967 Act for women resident in England and Wales. Pregnancies which lead to miscarriages are not included. Maternities which result in one or more live or stillbirths are counted once only.
- A.2.3 For live births, the date of conception is estimated as 38 weeks before the date of birth as gestation is not collected at birth registration. However, for abortions and stillbirths, date of conception is calculated using information available on gestation. In summary, date of conception is calculated as follows:
- For live births: date of birth – 38 weeks
 - For stillbirths: date of birth – gestation weeks
 - For terminations: date of termination – (gestation weeks + 2 weeks)
- A.2.4 ONS publish conceptions data annually by usual place of residence and age of the mother at conception. Information on the detail provided at different geographical levels in the publication is listed below:
- National - counts and rates for girls aged under 20 by single year of age and then by 5 year age-groups including percentage of conceptions leading to abortions. In addition rates for outcome (maternity or abortion) are provided annually for under 18s and as a 3 year aggregate for under 16s.
 - SHA/regions in Wales – counts and rates for girls aged under 18 and then by 5-year age groups including percentage of conceptions leading to abortions. Rates for outcome are provided annually for under 18s and as a 3 year aggregate for under 16s.
 - County/Unitary Authority (UA) – counts and rates for under 16s and under 18s and then by 5 year age-groups. Percentage leading to abortions are also provided for the under 18s.
 - Local Authority (LA) – counts and rates for under 18s and percentage of conceptions leading to abortions. Rates for outcome are provided as a 3 year aggregate for under 18s and under 16s.
 - Ward – total count and counts for under 18s
- A.2.5 The relationship between the abortions, conceptions and birth registrations needs to be examined in order to establish the risks for abortions data associated with releases of conceptions.
- A.2.6 Consider conceptions data for 2003. It is impossible to calculate the true count of abortions in 2003 as one would need to subtract from the conceptions data for 2003 the number of live and stillbirths that occurred in 2003 and also the proportion that occurred in 2004. Then one would also need to know conceptions that occurred in 2003 but were aborted in 2004. Finally one would need to add conceptions that occurred in 2002 but were aborted in 2003.
- A.2.7 Data by this breakdown are not publicly available and if an intruder did attempt to model/estimate the abortion figures the result is highly unlikely to be accurate and the time effort and expertise required is considered to be disproportionate. Some ambiguity will also be added to the numbers due to multiple births only being counted once in the conceptions data and because birth statistics are based on the age of mother at birth, rather than age at conception.

- A.2.8 ONS publishes live births by age of mother at birth for different age breakdowns at the National, SHA/region and County level. At the LA level live births are available by some age breakdowns including <18. At the ward level live births are only published as a total, no age breakdown is provided. Stillbirth information is published at National and SHA/region levels and disclosure controlled data for areas below SHA/region level is made available on request.
- A.2.9 In order to make any disclosive inferences about abortions from counts or rates of conceptions at National, SHA/region, County/UA or LA level an intruder will need to have access to counts of maternities (live + still) aggregated by the same age bands and for the same geographical level as for the conceptions data. Due to the time lag between conceptions data and births/abortions data, the unknown number of multiple births and the differences in the age of the mother at birth and at conception, the count obtained by subtracting maternities from conceptions is highly unlikely to reflect the true number of abortions.
- A.2.10 At the SHA/region, County/UA and LA level data on the percentage conceptions leading to abortions is provided. Although the percentages are rounded to the nearest integer in many cases the original count of abortions can be estimated. Although protection against disclosure of the number of abortions will be provided by the unknown number of stillbirths, multiple births, differences in age at conception and birth, and the time lag between conception and birth/termination the effort and expertise required by the intruder to obtain an estimate is not great. Therefore more protection is required. If the count of abortions estimated from the percentage conceptions leading to abortions is considered unsafe (e.g. count <10) then in order to be consistent with the abortion statistics such cells will need to be disguised in the conceptions publication.
- A.2.11 Rates of conceptions leading to abortions or outcome rates are published at the National, SHA/region and LA level. The population at risk denominator is available from the ONS mid-year population estimates. If the denominator is large enough and the rate is rounded then an intruder cannot recover the original count precisely. However in other cases an intruder could calculate a count of abortions from these rates that is disclosive. Therefore in the same way as percentage of conceptions leading to abortions, these counts need to be protected and suppressed if the underlying number of abortions is unsafe (e.g. <10 for any geography below the Government Office Region in England or the country of Wales and <5 otherwise).
- A.2.12 Abortion statistics are published by SHA/region and PCO/LHB geographies whereas conception statistics are published by SHA/region and LAs. PCO/LHB are non-coterminous with LA. An intruder could calculate an estimate of an abortion count for a LA from a rate or percentage of conceptions leading to abortion. The intruder could then difference the result with a published abortion statistic for a PCO/LHB to obtain an estimate for a smaller geography. Some further protection of the LA outcome rate or percentage conceptions leading to abortions figures may be required to cover this case of disclosure by differencing. Analysis should be undertaken to investigate the differences between the two geographies to highlight any problem areas.
- A.2.13 At ward level no data are available on stillbirths and live births are only reported for all ages. An intruder could difference the total number of conceptions and the total number of live births to arrive at an estimate of the total number of abortions + stillbirths for the ward. Protection against disclosure of the number of abortions will be provided by the unknown number of stillbirths, multiple births and the time lag between conception and birth/abortion. The number of abortions for under 18s cannot be derived from the conception statistics since data on stillbirths and live births is not available for this age band.
- A.2.14 A study was undertaken by the ONS to evaluate whether information available locally on the number of births for under 18s could be used in conjunction with the conceptions statistics to identify the actual number of abortions performed in a ward. The number of

conceptions and births to girls aged under 18 in 2000 and 2001 by ward boundaries were used in the analysis. Using the abortion data provided by DH the number of abortions performed to teenage girls in 2000 and 2001 by ward were produced. The difference between conceptions and births by ward was calculated to produce crudely conceptions leading to abortions (crude since multiple births and the time lag were not taken into account). These were then compared with the actual number of abortions by ward. In 2000 and 2001 the number of conceptions leading to abortions calculated crudely were the same as the actual number of abortions in 7% of wards in England and Wales. This reflects the protection that is provided by the unknown number of multiple births and the time lag between conception and birth/abortion.

A.2.15 The conclusion is drawn that conception statistics displayed as counts or rates at National, SHA/region, County/UA, LA and ward cannot be used to disclose information on abortions. Therefore counts and rates should continue to be published using current disclosure control methods. Where figures are published for outcome rates or percentage conceptions leading to abortions at National, SHA/region, County/UA and LA level protection will be required where the derived number of terminations is considered unsafe (e.g. <5 or <10). Some further protection of the LA outcome rate and percentage conceptions leading to abortions may be required to cover the case of disclosure by differencing.

A.2.16 Any changes to the data available on births, stillbirths or conceptions would require a detailed risk analysis.

Appendix 3 – Disclosure Control Methods

A.3.1 This appendix contains a technical description of each disclosure control method outlined in Chapter 5.

Table Redesign

A.3.2 For a particular variable unsafe cells can be disguised by collapsing categories. There are three broad types of table redesign:

- Top or bottom coding - where values at the very top or bottom ends of the distributions of continuous variables are recoded into single categories (e.g. age under 15 years; over 20 weeks gestation);
- Broad-banding continuous variables so that a response is recoded to lie within a particular range of values (e.g. using age bands of 16-25 years, 26-35 years, etc.);
- Broad-banding or collapsing categorical variables so that they are grouped together into one category (e.g. combining similar medical procedures).

A.3.3 Generally, the choice of which categories to combine would depend on several factors. Those categories with unsafe cells should be selected, and combined where possible with "similar" categories. Also, two smaller similar categories might be combined to form a larger one, but if they are dissimilar each should be combined with a different larger category to minimise the relative data damage. It is important to take into account how the proposed change will affect the consistency between tables and historic comparisons.

Example

A.3.4 An age variable may have been categorised into age bands such that cross-tabulation with a second spanning variable (type 1 or type 2) produces the frequency table below (Table 1). In this case the shaded cell contains a frequency of 1. As a result, the Age <12 column is potentially disclosive.

Table 1: Age and Type

	Age				
Outcome	< 12	12-15	16-19	> 19	Total
Type 1	1	5	7	6	19
Type 2	7	15	18	19	59
Total	8	20	25	25	78

Redesigning the age ranges can remove this potential disclosure. In Table 2 the lowest age band has been broadened and the frequency of 1 has become a 6.

Table 2: Age and Type

	Age			
Outcome	< 16	16-19	> 19	Total
Type 1	6	7	6	19
Type 2	22	18	19	59
Total	28	25	25	78

A.3.5 Note, collapsing categories does not necessarily have to be implemented across a whole table but can be applied to sub-tables, e.g. if a larger table existed that included the information above with other variables broken down by age then it may not be necessary to collapse the age bands for the whole table but just the outcome shown in the example

Combining Years

A.3.6 A safe way to increase access to more detailed counts is to publish three (say) year aggregates without any overlap. Not only does this increase the level of data that can be output, it also adds protection because the data is uncertain in timing (between 1 and 3 years). In addition it becomes much more difficult to make an identification due to the time lag and migration issues. Self-identification is still an issue, but if the individual has moved, then maybe it would not be so distressing/damaging. The rules for defining unsafe cells are the same as those used for annual data.

A.3.7 An alternative to publishing aggregated data without any overlap is to publish rolling aggregates, e.g. 2001+2002+2003 and then 2002+2003+2004 etc. If rolling aggregates are to be implemented then the rules for defining unsafe cells are not always straight forward and needs careful consideration especially if any year of the rolling aggregate has been previously published.

Record Swapping

A.3.8 Pretabulation techniques focus on perturbations of individual records using either a targeted or a random selection process. One such method is record swapping. This involves swapping pairs of records that are partially matched (e.g. individuals that have the same age) which alters the geographic locations attached to the records, but leaves all other aspects unchanged. The effect on tabulations produced from the record-swapped data is that some of the data will be counted in the table for a different geographical location, depending on the level of geography chosen.

Suppression

A.3.9 A method of protecting unsafe cells in tables is cell suppression. This means that unsafe cells are not published - they are suppressed and replaced by a special character, such as '.' or 'X', to indicate a suppressed value. This should be a different symbol to that used for missing values. Such suppressions are called primary suppressions. To make sure the primary suppressions cannot be derived by subtractions from published marginal totals, additional cells are selected for secondary suppression. The selection of secondary suppressions can be done either by hand or by software.

Example:

If a threshold rule of < 5 were applied to Table 1 the frequency of 1 in the shaded cell in Table 1 could be replaced with an X, as in Table 3.

Table 3: Age and Type

	Age				
Outcome	< 12	12-15	16-19	> 19	Total
Type 1	X	5	7	6	19
Type 2	7	15	18	19	59
Total	8	20	25	25	78

The suppressed cell in Table 3 can be derived by subtractions from the marginal totals. It is therefore necessary to carry out secondary suppressions.

Table 4: Age and Type

	Age				
Outcome	< 12	12-15	16-19	> 19	Total
Type 1	X	X	7	6	19
Type 2	X	X	18	19	59
Total	8	20	25	25	78

Table 4 shows the secondary suppressions (in the shaded cells) that are needed to ensure that the primary suppression is effective. Secondary suppressions should be chosen to minimise information loss e.g. select internal cells before marginal totals and smaller counts before larger counts. Care should also be taken to ensure that suppressions are consistent throughout all releases. The process of secondary suppressions can become very complex if the number of suppressions is not very small and software (e.g. the Tau-Argus disclosure control tool (<http://neon.vb.cbs.nl/casc/>)) should be implemented in order to ensure a safe and optimal solution.

Rounding

A.3.10 Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell while adding a small but acceptable amount of distortion to the data. Two alternative rounding methods are outlined below: random rounding and controlled rounding. In each case there is a choice of the base for rounding - common choices are 3 and 5. All rounded values (other than zeros) are then integer multiples of 3 or 5, respectively.

A.3.11 Note, conventional rounding (where each cell is rounded to the nearest multiple of the base) is not recommended here since the method does not offer suitable protection.

Random rounding

A.3.12 In random rounding, each cell value is rounded in a random manner, independently of other cells, usually (although not always) to an adjacent multiple of the rounding base. For example, values of 6, 7, 8, or 9 could be rounded to either 5 or 10 based on assigned probabilities. Various probability schemes are possible, but an important characteristic is that they should be unbiased i.e. there should be no net tendency to round up or down.

Table 6 shows Table 1 randomly rounded to base 5.

Table 6: Age and Type

	Age				
Outcome	< 12	12-15	16-19	> 19	Total
Type 1	0	5	5	5	20
Type 2	10	15	15	20	60
Total	5	20	25	25	75

With random rounding, there may again be inconsistencies in the data within tables (non-additivity, e.g. row 1 does not sum to 20) and between tables (the same frequency being rounded to different values in different tables).

Controlled rounding

A.3.13 In controlled rounding, values in the cells of a table are rounded to a multiple of a common base in such a way as to preserve additivity to subtotals and table totals. The controlled rounding method works for hierarchical data and for linked tables. Table 7 shows a possible controlled rounding solution for Table 1.

Table 7: Age and Type

	Age				
Outcome	< 12	12-15	16-19	> 19	Total
Type 1	0	5	5	5	15

Type 2	5	15	20	20	60
Total	5	20	25	25	75

Controlled rounding method has to be implemented in a statistical disclosure control tool, e.g. the Tau-Argus software (ref).

Barnadisation

A.3.14 Barnadisation is a form of random record swapping for frequency tables. The procedure modifies each internal cell of every table by +1, 0 or -1 according to the probabilities ($p/2$, $1-p$, $p/2$). Zeros are unadjusted. The totals are added up from the perturbed internal cells. Typically, the probability p is quite small and therefore the majority of cells are not modified. As in most post-tabular adjustments it leaves inconsistent totals between tables. In addition, as in random record swapping, it leaves high risk in the small cells, i.e. the probability that a one is a true one is quite high. If the true value was a one, then with probability $p/2$ the value is perturbed to a zero, with probability $p/2$ the value is perturbed to a two and probability $(1-p)$ the one remains a one.

