Risk of Disclosure – Diagnostics Monthly Return

Coverage

This paper assesses confidentiality and data disclosure issues with both inpatient and outpatient parts of the monthly diagnostics (DM01) return, for both providers and commissioners.

Background

1. Statisticians have a professional duty to protect the confidentiality of individual level data obtained to produce statistics. The Code of Practice for Official Statistics sets this out in Principle 5: "Private information about individual persons (including bodies corporate) compiled in the production of official statistics is confidential and should be used for statistical purposes only". The Code of Practice also states arrangements for confidentiality protection should be sufficient to protect privacy but not so restrictive as to limit unduly the practical utility of statistics. The main legal instruments governing this balance are the Data Protection Act, which places obligations on organisations to protect personal information and the Freedom of Information Act, which creates a public right of access to information.

2. The design of a statistic should meet the obligation to protect against disclosure, but should then be optimised to include as much detail in the statistic as reasonably possible, to fully meet the needs of the users.

3. There is a need to assess whether this data is potentially disclosive.

Guidance from ONS – the structure of this assessment

4. Guidance from ONS[1] on confidentiality sets out guidelines for any assessment of disclosure risk. It stops short of setting out hard and fast rules, but is clear on the need to protect patient confidentiality while at the same time maximising public access to official data. This guidance summarises the six main steps for ensuring access to non-disclosive statistics as shown in Figure 1.

[1] GSS/GSR Disclosure Control policy for tables produced from administrative data sources (the document is available at the following link: http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-tables/index.html)

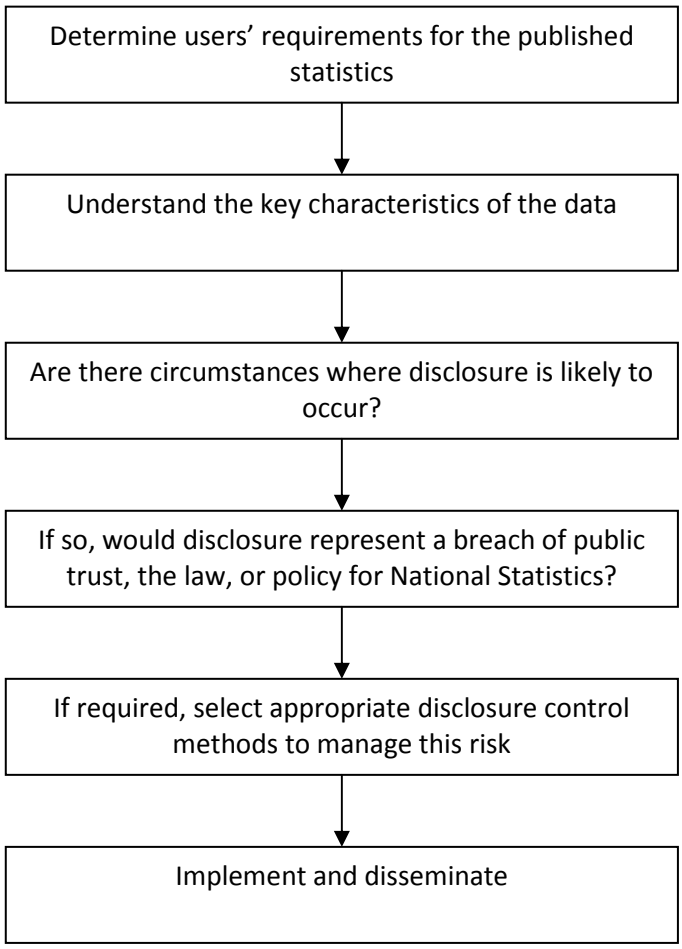

Figure 1: Main steps for ensuring access to non-disclosive statistics

Step 1 – Determining users' requirements

5. The requirements for this data were set out in the monthly diagnostics definitions published in 2006 and last revised in 2009. This document includes the purpose of the collection and highlights its intended uses.

6. A collection was established using UNIFY2, to collect this data on a monthly basis from all English providers, and the corresponding organisations that commission them.

7. The current collection allows members of the public and those working within the system to have access to up-to-date information. This leads to an implied need to publish data subject to any confidentiality constraints in a timely way.

8. The monthly DM01 data is published to give patients an insight into the waiting times and activity levels of their local trust or commissioner, and allows them to compare against all other trusts and commissioners in England. The areas covered in DM01 include:

- Waiting times for patients waiting for one of 15 key diagnostic tests, split by number of weeks waiting
- Activity in the month on each of the 15 tests

9. There is converse public interest in ensuring that information about the experience of individuals is safeguarded in an appropriate way. A balance must be struck between measures to protect confidentiality and the public good arising from publication.

Step 2 – The characteristics of the data

10. This is an aggregated data source. The data is submitted by providers based on patient level information that is taken from an administrative data source within the trust.

11. There is a process of data cleaning and validation within the collection system. This allows for periodic revisions, twice a year, should trusts identify that they have more accurate data once the data have been published. Prior to the publication, the central data team undertakes some basic checks to identify if data is significantly out of step with other trusts and previous submissions.

12. The majority of the data collection does not present a risk of disclosure, but depending on trust/commissioner size and general activity levels, a number of indicators could potentially return small numbers. These include waiting times for the longer timebands (eg 6 plus weeks) for commissioner data and all indicators for provider based data. This is particularly the case for small trusts such as community trusts.

Step 3 – Evidence of risk of disclosure

13. Publication of any detailed data may increase risks of disclosure of information relating to an individual patient. It is important to note that these data do not include any personal identifiers, so it is not possible to identify patients directly from the published data. Instead the categories of disclosure risk (situations in which disclosure might arise) are as follows:
    - Self-identification risk: When a patient recalls their circumstances during the time-period of the data collection and can recognise, from the context, which data refers to them. A patient sees a 'one' in a table and is able to recognise from context that they are the 'one'. Following cases taken to the High Court while in law applies to those individuals who can identify themselves within a larger count, so the "need to be sure that publication would not cause, or be likely to cause, unwarranted and substantial damage or distress" in theory applies. However, in practice it appears that the damage is only likely to be considered substantial if there is a rational fear of being identified. Even in high profile cases it has been deemed that the potential disclosee fear was irrational and so there was no breach of confidentiality.
    - Motivated intruder risk: Where there are reasons for a third party to seek further information about cases of a patient, for example where a 'celebrity' case arises or where cases in a particular organisation happen with a newsworthy frequency or pattern. This type of risk can be broken down further into two types:

a. Identity disclosure: Where a third party is able to determine who the data relates to using the data itself and other information available to that third party.
b. Attribute disclosure: Where a third party is able to infer additional information about an individual.

It can be concluded that there is no risk of identity disclosure, as the possible population size of the collection is large and the collection does not contain any personal identifiers. Instead, the paper will focus on the motivated intruder risk in regards to attribute disclosure.

Self Identification risk

14. There may be circumstances where a patient can self-identify. Current published tables can contain small numbers. This is not in itself a reason for suppressing data. An appropriate test is defined by the Data Protection Act 1998, which requires the matter to be considered (although it does not directly require all self-identification to be avoided). There is a need to confirm that the published data would not cause, or be likely to cause, unwarranted and substantial damage or distress.

15. As the collection is related to waiting times data, self-identification is likely as it only requires recognition of hospital experiences during the time period. This identification is similar for both commissioner and provider.

16. It is highly unlikely that distress would be caused by self-identification unless some sort of negative emotion is evoked from recognising the patient's hospital activity, or, for example, that the patient had waited a certain number of weeks.

17. The patient would recognise that only someone who already knew about their activities and location, availability during organised appointments and their area of residence would be able to identify them, and therefore no additional information is revealed.

18. If someone had access to PAS[2] or HES[3] data then identification could be possible, but both these data sources are subject to their own security and rules concerning confidentiality.

19. The broad conclusion is that there may be a risk of self-identification, however the consequences of this are highly unlikely to cause damage or distress to the individual patient. There is therefore no need to suppress any small numbers to avoid self-identification.

20. Thus it can be concluded that whilst some self-identification may be possible, it is highly unlikely to cause any damager or distress to any individual.

[2] A Patient administration system (PAS) is core component of a hospital's IT infrastructure. It records the patient's demographics and details all patient contact with the hospital, both outpatient and inpatient.

[3] Hospital Episode Statistics (HES) is a data warehouse containing details of all admissions to NHS hospitals in England. More information is available here: http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937

Motivated intruder risk

21. The risks of being identified by a third party are similar to those arising from self identification, expect in the following aspects:
    - The third party may not have access to information that the individual is aware of (regarding themselves), so in some areas risk is reduced.
    - However, it may be a breach of confidentiality if a third party can deduce anything about the individual.
    - We need to consider carefully the extent to which a third party might become a motivated intruder, with an incentive to explore the data and deduce information about the individual.

22. The published data does not contain any personal identifiers. The additional risk that publishing small numbers allows a motivated intruder to deduce information about an individual is next considered.

23. The incentive, and consequently the risk, may be higher when celebrities are known to have attended hospital during the quarter. There may also be scenarios where someone would seek information about a friend or relative.

24. For example, assume in one month a trust's only patient waiting was a patient waiting for a MRI scan. If that scan was reported as patient waiting over say 13 weeks, then it is possible for a third party to infer that the patient had been the one who had waited that long. The patient could only be personally identified if the third party already knew that the patient had been waiting for a scan at the month end.

25. It is impossible to infer a patient's identity from the activity data alone. Prior knowledge or access to personal information would be required.

Step 4 – Would disclosure represent a breach of public trust, the law, or policy for National Statistics?

26. Previous GSS protocols on confidentiality stated that disclosure control methods should be judged sufficient when, taking account of information likely to be available to third parties, it would take a disproportionate amount of time, effort or expertise for an intruder to identify a statistical unit to others, or to reveal information about that person that is not already in the public domain.

27. In this collection there is no additional data from which an individual can be identified. If a third party was able to access other data sources, such as HES or PAS, to further identify a patient, these secondary sources would in themselves have to be full disclosive in their own right in order for an individual to be identified. As discussed above, HES and PAS have their own security protocols.

28. Where patients can identify themselves in the data, there is a risk that the patient could view this as disclosive. As discussed above, this self-identification risk is not a substantial one. Disclosure would not represent a breach of public interest.

Conclusion

29. The risk of disclosure and/or harm or distress to data subjects is minimal. It is not possible to infer some other fact about the patient's condition or treatment using

this data set alone, nor in conjunction with information likely to be available to third parties.

30. It is possible that some patients will be able to identify themselves, but there have been no instances of public disquiet about this and risk of harm from self-identification is very low.

31. For this reason, the monthly diagnostic return should be considered minimal risk.