BRIEFING – Health Intelligence and Information Governance

> **"It is not justified to share information that identifies people, when anonymised or statistical information could be used as an alternative."**
>
> *ICO Framework Code of Practice for Sharing Personal Information, p8*

Covering:
- Privacy Breach Risk
- Methods of privacy attack
- PBRisk Measurement
- Privacy dilemma and the Evidence Base
- The PARiP Approach

P-Day is less than a year away

Don't be an April Fool

The ICO Code of Practice sets out key principles with increasing importance for everyone using patient data. From CEs to administrators, commissioners to researchers, the pressure is on to ***demonstrate*** that data is being handled in a privacy-friendly manner. Traditional practices will be under increasing scrutiny.

Across the NHS, the bar is raising for privacy-compliant handling of personal data. In April 2009, Pseudonymisation[1] Day or "P-Day" marks an important milestone in this process when ***The Operating Framework for 2008/09*** stipulates that the Secondary Uses Service (SUS) should be used as the standard ***pseudonymised*** patient data repository for activity for performance monitoring, reconciliation and payments. *(p35, section 3.35)*

> ***"Where there is a need to link data from different data sets or over time, linked pseudonymised data should be used..."***
> *Report of the Care Record Development Board on the Secondary Uses of Patient Information, p11, section 4.2.i*

Enforcing the principles of *Confidentiality: NHS Code of Practice*, P-Day is part of a trend ensuring that all re-purposed data be de-identified.

The current position across the health "industry" with regard to privacy compliance is very patchy. Many organisations look likely to be caught out with systems and procedures that will not pass muster. It is hardly an exaggeration to say that a data handling time-bomb is ticking away at the heart of the health service. It can only be de-fused if concerted action is taken promptly with a clear drive from the top. This paper suggests how.

[1] In non-technical terms, pseudonymisation is logically the same as the adoption of a "nom-de-plume" or "nom-de-guerre". In this context however, the "nom" is allocated to an individual by a complex mathematical process rather than chosen by them, and the individual will not know the pseudonym that has been allocated to his or her data.

SAPIOR
ENABLING ETHICAL DATA SHARING

Privacy matters - in health and elsewhere

Right across the public and private sectors, there is a groundswell of public concern about the use of personal data by organisations and a growing unease about potential uses which may damage the interests of individuals. From concern with the use of loyalty card data by supermarkets to the proposed National Identity Card Scheme, privacy has risen to the top of the public agenda. This should be no surprise to the medical community since confidentiality of the data that is their stock-in-trade is one of the oldest principles of medical ethics, and fundamental to patients' perception of medical staff and health service providers.

Perhaps what is happening here is the community at large is coming to believe that the principles of privacy which have been generally thought to apply to medical data should be applied to other data about them. This public concern is in its relatively early days but it is easy to see that issues of privacy have been emerging in a number of different domains over recent years and that resolution of these issues will be not be straightforward.

Privacy Breach Risk

In everyday discourse privacy could be viewed a black-and-white issue: our data is either completely private and can never be compromised or else is completely open to inspection by anyone who is interested. Of course, neither extreme is actually possible and it is more useful to consider privacy along a spectrum of Privacy Breach Risk (PBRisk) which reflects the probability that personally identifiable data will be seen by an unauthorised and possibly ill-intentioned person.

From this perspective, it is clear that risk can never be at either extreme. There is no such thing as complete privacy, nor is it likely that any health sector business process would ever publish private data!

The challenge of privacy therefore needs to be cast as one of minimising PBRisk in a given scenario rather than eliminating it altogether.

As the expression of PBRisk (see box) makes clear, its minimisation is a complex matter because it requires consideration of interacting issues of people's behaviour and susceptibilities, the technology of data processing and the business processes or governance of organisations.

Each of these dimensions of the challenge needs to be addressed in a systematic way. But, clearly, if data sets are routinely pseudonymised as they are copied for repurposing, the risk of privacy breach due to system failure or individual misbehaviour is radically reduced. This is because the data can no longer be inadvertently or accidentally identified, meaning that only deliberate misbehaviour could breach privacy.

Expressing Privacy Breach Risk (PBRisk)

The PBRisk may be seen as a mathematical expression, the terms of which are factors such as:

- Number of people with access to the system - data used by one researcher in a single location vs. by a multi-disciplinary team in several places
- Type of organisation with access to the data – contained entirely within the UK health industry vs. has a US parent body that could possibly be obliged to disclose information to US organisations under the Patriot Act
- Characteristics of people using the data - personal reputation as researchers which could be threatened by a breach, personal vulnerability to bribery or blackmail to release private data, likelihood that a deliberately infiltrated criminal could access data, etc.
- Intrinsic interest of the data - certain kinds of health data (e.g. on public figures or celebrities) will be of greater prurient interest than others and thus more at risk
- Integrity of role-based access systems - design of systems intended to ensure that data can only be accessed by people with the correct authority and need to know
- Policing and enforcement of processes - extent to which operational procedures designed to protect privacy are enforced in practice and regularly audited for compliance
- Scale of sanctions for deliberate malfeasance - severity of penalties for proven misbehaviour
- Nature and purpose of analysis permitted on a data set - extent to which analysis is supervised by authorities able to prohibit unethical or privacy-threatening activity
- Level of data "de-identification" - sophistication of data transformation applied; ranging from "perimeter security" only (protecting machine where data resides but is clearly readable) to data pseudonymised by a multi-stage process built to DH specification (rendering personalised data unreadable even if the perimeter is breached)

Methods of privacy attack

Of the terms expressing PBRisk, the "nature and purpose of analysis permitted" raises another dimension of privacy attack that could be mounted on a data set even if it has been comprehensively pseudonymised.

This is the so-called "inference attack", where it may be possible to infer the identity or attributes of an individual within a large data set by analysing interrelated characteristics to progressively narrow down possible entities in pseudonymised data such that an individual can in fact be identified.

This kind of attack, of course, is much more resource hungry than an attempt to break the encryption of a data set and will typically be attractive only where the person trying to breach privacy is hunting for a particular target, such as a celebrity of some kind. While this kind of breach is of course a serious matter for the individual concerned, it is not of the same significance as a wholesale breach of privacy for a large number of individuals by gaining access to an entire data set. The inference attack route therefore—while it remains open even after a data set is pseudonymised—is therefore a relatively small component of the overall PBRisk.

Faced with this level of complexity, there is a danger that policy will become a victim of paralysis through analysis. So it is always necessary to make pragmatic judgements about the level of protection proportionate to the damage that could arise from a breach. The terms in the PBRisk expression, however, set an agenda that must be coherently addressed. If this is not done, it is equally clear that the time-bomb could yet do immense damage to the image of the health service.

This is why P-Day is so important. It sets a date by which a fundamental contribution to the minimisation of PBRisk should be implemented. By pseudonymising data at the individual patient level, measures to address physical security of data processing or staff integrity have a kind of long-stop ensuring that if all else fails, the data itself

"At present, many users seem either unaware of the details of the policy or choose to ignore it, and mechanisms to enforce the Confidentiality guidelines are not in place.
This has the result that there is a gap between policy and practice.
Secondary data users are gaining access to clear data for purposes they feel are legitimate and generally there is no comeback for breaching the policy unless an egregious breach has occurred."
Secondary Uses Service Pseudonymisation Impact Assessment Study, p9, section 2.2.6

is protected. Without pseudonymisation, sensitive data is only protected with the strength of the weakest link in the chain of security measures surrounding it.

"Information at the centre of health"
Privacy Enhancing Technologies and business processes may seem rather remote from the day to day concerns of delivering health care. In fact, their significance is increasing in step with the growing recognition that "putting information at the centre of health", as expressed in the DH Public Health Information Strategy, is vital to both quality of care and efficiency of operation —the twin concerns of public debate about health provision.

So far as medical practice is concerned, the availability of more detailed information and the use of more sophisticated techniques open up possibilities for more penetrating analysis which will pay off in improved clinical outcomes.

For the developing commercial environment of "world class commissioning", the importance of clinical data is increasing. The analysis of need and service provision within a geographical area is fundamental to the business cases that Monitor expects to evaluate in authorising and subsequently checking performance of Foundation Trusts. And the entire concept of "Payment by Results" is predicated upon the ability to analyse data to demonstrate those results and their responsiveness to clinical need.

In health and more widely across the public sector, the need for data to support the increasing use of "Lean" analysis of business processes has also created an environment in which it is vital to have a command of information.

The use of Lean techniques in manufacturing has revolutionised industry over the last five decades. The public sector is now learning to apply these insights to its processes. There is, however, a critical difference between these domains: the data on the performance of inanimate materials and of employees is naturally available for analysis without any constraint. In the case of medical matters, the raw materials of health processes are individuals whose data belongs to them—not to the institution—and is highly sensitive.

Privacy dilemma and the evidence base
This consideration poses the apparent dilemma of privacy in health. On the one hand, the management of health interventions and of institutions can be improved by analysis of data. On the other hand, the data in question deals with some of the most sensitive personal information: data about which surveys have shown that people are most concerned.

This dilemma is also inherent in the Prime Minister's recent espousal of a "personalised and empowered service". *(HSJ, 3 July 2008)*

"Personalisation"—especially where it requires tracking of individuals over time to judge the cumulative effect of health interventions—demands a robust approach to handling data. Similarly, "empowerment" demands that stakeholders or actors in the health sector need to be able to access and analyse data to help them improve service, without endangering the ethical values of patient confidentiality.

As discussion between professionals on all aspects of medical practice has increasingly emphasised the need for "evidence-based" approaches, an increasing load has been placed upon the collection and analysis of clinical data. So the tension, as referred to in the introduction, has grown.

Faced with conflicting pressures to use information to improve service and the dictates of medical ethics, the result has—frankly—been a fudge. As was pointed out in the *Pseudonymisation Impact Assessment Study*, which set the aspiration for "P-Day" in 2009, the result is "a gap between policy and practice" with "no comeback for breaching the policy unless an egregious breach has occurred."

The move to use SUS will work for access to national data sets. But SUS won't meet all de-identification needs, especially where organisations have multiple data sets, including locally held data, needing to be joined.

Fortunately, this tension can be relaxed, and the gap bridged, if a radically different approach is taken to the handling of personal health data.

This new approach is partly achieved through technology as will be described below. However, there also needs to be a step change in the manner by which personal information is handled, as noted by Walport and Thomas when reporting on their Data Sharing Review.

"We must all play our part in bringing about the cultural transformation that is needed if we are to secure the benefits, but minimise the risks, of sharing personal information."
Covering note for publication of the Data Sharing Review Report by Richard Thomas and Mark Walport

Relax the tension, bridge the gap
The privacy dilemma is only present where personally identifiable data is repurposed. If it were possible to ensure that all personal data was de-identified at the point of origin but capable of being re-identified as needed at some later point, the PBRisk would be dramatically reduced. Users of re-purposed data could access the full "texture" of the data to gain insight and understanding without compromising patient confidentiality.

This is exactly what the Sapior approach to data management achieves. It does this through routine pseudonymisation of data at the point of copying it to a system in such a way as to allow re-identification downstream. Of course, re-identification is only possible where the need to do so can be demonstrated and permission is given through a well defined business process.

This is the PARiP approach: Pseudonymise All, Re-identify if Permitted.

The PARiP approach

Many of the problems of data confidentiality occur in practice because the re-purposed analysis, for which data will be used, cannot be specified in advance. This is the nature of scientific analysis: to inevitably pose new questions in light of the answers to the researcher's original questions. The same is true for the kind of analysis which lies behind the commercial questions and issues that will be addressed as "world class commissioning" proceeds.

Approaches to privacy which destroy the texture of health data—such as small cell or "outlier" data suppression or aggregation—will inevitably impede research. Pseudonymisation avoids this problem. Inference attacks are still possible for those who would wish to probe data with a malign purpose, but are effectively blocked by appropriate business processes—such as staff screening and monitoring—and analysis protocols (where the management of these processes should be proportionate to the damage that could be inflicted by mischievous use of data).

Privacy in the use of health data is a multidimensional problem to which there are no easy answers. Nor, indeed, is there any system which can absolutely guarantee confidentiality. Such perfection is not given to human kind!

Sapior Pseudonymisation:

- Allows for re-identification as needed if permitted
- Patent-pending multi-stage technology
- Speed of loading is not significantly affected
- Highly scalable—data volumes increasing over time handled smoothly
- Sapior "black box"—a specialised hardware component which is compatible with a Service Oriented Architecture (SOA)
- Licence and maintenance costs geared to usage
- Provides DH SUS standard pseudonymisation for your data centre

What can be achieved is an unpacking of this complex problem into a number of significant dimensions which allow a balanced and proportionate approach to privacy to be taken. Such an approach recognises that privacy cannot be absolutely guaranteed, but is based on understanding the trade-offs and compromises that must be made between conflicting requirements.

The routine de-identification of data being copied/loaded into a data warehouse or analysis system creates a new paradigm for the handling of data in both patient-focused and management system-focused uses without affecting the use of patient identifiable data for the original clinical purpose.

PARiP= Pseudonymise All, Re-identify if Permitted

PARiP reduces the ability and opportunity for privacy breaches to be committed

At present, it is inappropriate to describe breaches of data handling principles in the health service as criminal. Although it is not hard to imagine some circumstances in which this would be the case! But it is instructive to borrow some insights from forensic examination of crime to throw light onto this issue. The conventional forensic approach to appraise the risk of crime being committed is to consider four key questions:

- Which people might be tempted?
- What ability do they have to commit crime?
- What motive do they have?
- What opportunity do they have?

In the context of the health service, clearly PARiP addresses two key dimensions—ability and opportunity—by making the illicit use of data very much more difficult. The major benefit of PARiP comes from simplifying and automating risk reduction.

How much will PARiP cost?

Following the PARiP approach, once data is de-identified whatever analysis is required for medical or administrative purposes can occur in many places and in depth without threatening the confidentiality of personal information.

Moreover, should such analysis identify a subset of the individuals described pseudonymously who would benefit from further medical intervention, the reversibility of pseudonymisation allows re-identification to be done in a controlled manner.

The Sapior module employs patent-pending multi-stage technology which ensures that the speed of loading is not significantly affected. Sapior's approach is highly scalable so that volumes of data increasing over time are handled smoothly. Sapior reidentifies individual records, once again, speedily and scalable to increasing levels of demand.

As mentioned earlier, re-identification must be under explicit permission control with a built in audit trail. This ensures that the request for re-identification meets explicit criteria designed to protect privacy and provides accountability for each instance of re-identification.

Taking both the initial pseudonymisation and the potential for re-identification together, the overall cost implication of adopting PARiP is likely to add around 2% to the aggregate running costs of the data processing operation at the PCT level. This of course is not trivial, but must be compared with two important factors:

- benefit from more readily available re-purposed data, without need for time-consuming, ad-hoc checking of compliance with confidentiality principles, and
- risk minimisation of an "egregious breach" which could entail massive costs in litigation and reputational damage.

What next?

P-Day is coming. And, with it, the enforcement of *Confidentiality*.

While it is possible, as *PIAS* reported earlier, there will continue to be "no comeback for breaching the policy unless an egregious breach has occurred", the key questions for CEs and research leaders to consider are:

- **Can this fudged approach continue in a post-Darzi world when the ICO has made it clear that current practice is "not acceptable"?**
- **How "egregious" does a breach have to be before a Chief Executive's head rolls?**
- **Who wants to be the first April Fool?**

www.sapior.com