



Preparing clinical study reports for external sharing

how to balance patient privacy/
data utility priorities and manage risk

Cathal Gallagher
d-Wise, Inc., Manchester, UK

Correspondence to:

Cathal Gallagher
d-Wise
Suite 7A, Manchester One
53 Portland Street
Manchester, UK M1 3LD
+44 7786 967185
Cathal.Gallagher@d-wise.com

Abstract

As the EMA refines its requirements for the external publishing of clinical study reports, the workload of medical writing teams is increasing to include robust processes for clinical study report anonymisation. Until now, life sciences firms have played this safe by using heavy content redaction (covering up identifying information with a blue box), but now EMA is encouraging anonymisation over redaction to help maximise data utility while simultaneously mitigating the risk of patient identification. (Anonymisation involves changing identifiers, but they are still readable, such as placing an age of 27 into a band of 20–29). This article explores the issues and considers companies' options.

EU measures to make clinical trial data open for public access have created substantial additional work for medical writers and transparency departments. In line with general shift towards greater transparency, companies must now tread a careful line between maximising the utility of clinical trial information and safeguarding patient identities as study reports are shared more widely.

Under EMA Policy 0070 on the publication of clinical study reports (CSRs) relating to medicinal products for human use, CSRs must be anonymised to prevent patients (and indeed professionals) who participated in clinical trials from being identified. The standard approach has been to redact anything that might identify an individual by using Adobe Acrobat software to cover that text with a blue box bearing the letters PPD for “protection of personal data”. In a supporting anonymisation report, the writing team explains what they have covered up and why.

There are several problems with this approach to patient risk management. The first is that it places a significant additional administrative burden on medical writers. Anonymisation via manual redaction is a labour-intensive process, taking up a considerable amount of trained experts' valuable time – given that CSRs can run from between 5000 to 100,000 pages – and sometimes even more. Second, it carries a risk. If just one potential subject identifier is missed, it would be quite easy for someone to piece together more specifics about the study – details which, in line with EU requirements, should not be disclosed outside of the immediate R&D team. A third significant issue is the impact of heavy redaction on the residual value of the amended content to interested external parties. If the goal is to make CSRs more open and available for external scrutiny, that aim is immediately compromised as soon as large sections of those reports are covered.

Smart approaches to identity safeguarding

It is this issue of clinical trials' external utility that has prompted new efforts by the EMA to dissuade life sciences R&D organisations from relying on redaction as their method of choice for report anonymisation. Instead of a very conservative “cover all” approach, EMA advocates that companies anonymise externally facing reports by using anonymisation techniques that can be adapted according to the perceived level of risk of patients being re-identified.

Using techniques such as date offsetting (assigning a random number to a patient and then changing all the dates related to that patient by this number) and other systematic (and internally traceable) alterations to identifiers, companies can confidently disguise revealing information while retaining the integrity of the findings and the surrounding narratives. An added benefit is that if an occasional identifier is missed, there would be nothing to suggest to the reader that it was a real clue regarding the original data; effectively, it would be hiding in plain sight. As a result, there is much less risk with this approach to the safeguarding of patient privacy.

Improving data utility through more accurate risk measurement

EMA has defined the acceptable risk level for



patient re-identification to be 0.09 – meaning that each subject's defining characteristics (country of residence, race, etc.) must be in common with those of at least 11 other patients taking part in the trial. One option if this is not the case is to anonymise data in a way that creates larger groups or equivalent classes – e.g., using “European” in place of “Irish”, or “other” for non-white ethnicity in a group with too few black or Asian subjects. Another option is to include subjects from other trials within the same therapeutic area within the same geographic area. This involves creating a larger population from which you are going to calculate your risk metrics. For example, if a sponsor is conducting several cancer trials within a given period, that information can be leveraged to help create a larger population on which you calculate risk. This is common practice when anonymising small trials.

The great advantage of this type of systematic approach is that information technology systems can take over much of the process, requiring only quality assurance checks from medical writing teams. Busy professionals are saved from doing all the legwork but they retain control over risk management. Systematic anonymisation is also much easier to audit internally, so teams can keep track of what they have done. They also will have a record of their actions, which they can use to demonstrate that all possible steps were taken to

protect the identity of patients.

One of the inhibitors to this kind of initiative has been a lack of drive from the EMA to make things happen, despite the agency's best intentions. To date, it has offered just guidance. Up to now, therefore, the majority of firms have continued to default to redaction, relying on outsourced services to fulfil the requirement if internal medical writing teams have not had the capacity. While not the most efficient and reliable approach, it has been seen as the least disruptive.

To continue in this vein is short-sighted, however. Other regions including North America and parts of Asia are already taking active steps towards anonymisation of clinical findings. Health Canada has already made in-roads with a very similar approach to EMA's, the FDA is likely to be next, and Japan is taking decisive steps too. The future will likely see a shift towards anonymisation, quantitative risk measurement and a focus on data utility. Whether guidance becomes law remains to be seen.

Going deeper: Anonymising underlying patient data sets

Although talk of the EMA extending its anonymisation requirements to individual patient data – i.e., underlying trial data sets – has not yet come to anything, it is an approach that offers maximum efficiency for the long term. Ben Rotz, director of medical transparency at Eli Lilly,

said at a recent conference that he did not expect Phase 2 of EMA Policy 0070 (data-level requirements) to be introduced in the next 10 years. But this does not take away from the process streamlining that is enabled by applying anonymisation techniques at the core. For one thing, it is the best way to get to an accurate risk score (enabling a quantitative rather than merely a qualitative risk measure). There are signs that some companies have recognised this, seeing the merits of linking documents to original data sets more dynamically.

Industry leaders are starting to apply more automated anonymisation methods to their CSRs, seeing the value in a more systematic approach. An added benefit is that associated anonymisation reports can be generated automatically, saving medical writing teams a lot of time and ensuring that nothing is left out in the explanatory notes.

These trailblazers are not doing this to score points but rather to reduce workload and to increase the consistency and value of their output. Although the EMA is still accepting any form of anonymisation, including redacted clinical reports, ultimately it is life sciences organisations that will suffer the consequences if a patient is re-identified because of inadequate risk processes.

Firms are now faced with maximising the utility of their data to external audiences while limiting the risk that individual patients will be identified. Systematic approaches to data-level anonymisation techniques offer the most flexible way to meet both goals. One of the outstanding issues to date has been that the EMA has not been very clear about the target audience for externally published report content. If it is the general public (e.g., interested patients), they are unlikely to understand the detail and language used in CSRs, so the value is questionable. If the audience is other researchers, it

could be argued that a summary and details of efficacy and adverse events would suffice. But interestingly, of the parties seen to access shared content to date, the largest audience has been other pharmaceutical companies – their main

driver for accessing the reports being to understand how their peers are approaching document anonymisation. Yet, as an educational resource on anonymisation, the current population of reports are not great examples of high-quality anonymisation combined with accurate risk metrics. There is plenty to improve upon.

Leading on data transparency: Pharma's time to shine

For now, progress depends on life sciences companies being able to see the bigger picture and appreciate the business benefits they can derive from being (a) more open and transparent with the market, and (b) more systematic and efficient in the way they manage personal data protection and risk.

Something to bear in mind is how quickly technology is developing and growing in sophistication, especially in the context of artificial intelligence (AI) and machine learning. Incorporating automated intelligence within anonymisation applications offers teams the ability to teach systems to more accurately recognise potentially sensitive references, links, and context – and to take specific action (e.g., offset all associated dates by X days), or flag them to the teams for checking. Intelligent software can pick up inconsistently spelled references, which a text search might miss.

Initial benchmarks suggest that by the seventh time a system has been shown something (e.g., what constitutes a sensitive identifier), it is already at human-level accuracy. After that, it soars ahead, becoming progressively better and faster. So productivity increases and leak rates (mentions being missed) drop significantly – to a level less than 1%. AI-based systems can also be set to apply different levels of risk mitigation – so if there are sensitivities about alcohol use or pregnancy, for example, anonymisation actions can be set accordingly.

But of course technology alone does not have all the answers. Organisationally, there are still communication gaps among medical writing,

privacy, and data teams, and between internal departments and outsourced service providers. These barriers, added to an unwillingness to take investments beyond the scope of basic compliance, will limit what companies are able to achieve – unless they proactively take steps to change things.

For patients' sake, it is essential that life sciences organisations are vigilant about protecting patient privacy and about regularly reviewing the risks of re-identification. In the interests of keeping pace with the way authorities' requirements are going, it is far better that companies move forward with higher goals now than remain behind the curve as the industry presses on with plans for greater transparency and collaboration.

An insightful observation made at an event recently was that while all sorts of companies, from banks to internet companies, are pulling out all the stops to collect data, very few are sharing it for the greater good. The pharma industry might be slow to adopt other technology trends, but it is taking a lead in data transparency. Make the right choices now and in 10 to 15 years' time companies could find themselves giving advice to businesses in other sectors about best practice strategies and describing how they got to where they are.

Disclaimer

The opinions expressed in this article are the author's own.

Conflicts of interest

The author is employed by d-Wise Technologies, a specialist provider of data and process optimisation services for life sciences and healthcare industries.

Author information

Cathal Gallagher is a senior life sciences consultant at d-Wise Technologies and an expert in data/CSR anonymisation. He is also an active member of the EMA's technical anonymisation group and of one of its subgroups whose remit is looking at new technological developments.

An insightful observation made at an event recently was that while all sorts of companies, from banks to internet companies, are pulling out all the stops to collect data, very few are sharing it for the greater good.