An interview with
Cathal Gallagher
of EMA's Technical Anonymisation Group

About this article

In our daily work in a company or freelance setting, we interact with other relevant functions, typically biostatistical, medical, programming, and data management colleagues, so that we can deliver well-rounded deliverables that take account of multiple perspectives. If we take a higher-level strategic view for the developing regulatory public disclosure (RPD) arena, then we should be talking to professionals outside of our usual networks to ensure we understand and take account of broader perspectives. I engaged with Cathal Gallagher of EMA's Technical Anonymisation Group and asked him the kinds of questions that you as medical writers might, and hope to have contextualised his RPD perspectives to our work in writing clinical documents fit for public disclosure.

Sam Hamilton

Sam Hamilton: I am delighted to be talking to Cathal Gallagher, a member of the EMA's Technical Anonymisation Group (TAG).[1] EMA's TAG is an expert group in data anonymisation, and they aim to help further develop best practices for the anonymisation of clinical reports, in the context of the EMA's policy on the publication of clinical data (Policy 0070).[2] Cathal, you have kindly agreed to tell us more about the TAG and what the work of the TAG means in the context of the work that regulatory medical writers do. Let's start by understanding a little about your career and how you ended up on EMA's TAG.

Cathal: I used to be an IT and Maths high school teacher. In late 2011, I decided that I needed a change and made an effort to retrain myself in SAS programming. SAS software is used to create the tables, figures, and listings (TFLs) that medical writers typically use as source data in their clinical study reports (CSRs). With my newly found skills, I was hired by a small company doing SDTM[3] and ADaM[4] programming. These two Clinical Data Interchange Standards Consortium (CDISC) standards[5] are used to develop clinical trial datasets and are the required standards for both FDA's and Japan's Pharmaceutical and Medical Devices Agency (PMDA) dataset submissions. This small company was bought by d-wise, who were building a piece of software to anonymise clinical trial datasets. I got involved in the project and we finished developing the software, which is called Blur. Meanwhile, we could hear industry rumblings about public sharing of clinical trial documents becoming a new area of interest. It seemed a logical – although not an easy next step – to expand the Blur software to meet these new needs. I found myself at every possible industry conference trying to learn about clinical trial data and document transparency, and I developed an understanding of the responsible clinical trial data sharing that is such an important consideration in the writing of clinical documents, for example, CSRs. I heard that EMA was setting up a TAG, and I applied for a role. I was pleasantly surprised when I was selected, and so here I am on EMA's TAG.

Sam: The TAG includes members from academia, industry, patients, and healthcare professionals. It's interesting that there are no medical writers on the TAG because we are largely responsible for making decisions about and anonymising texts within CSRs (and clinical summary documents) that may compromise the privacy of patients when those CSRs are made publicly available. We currently do this in a qualitative way, by proactively anonymising the text in our CSRs to protect individuals, whilst trying to maximise data utility. We know that our statistician colleagues are learning about quantitative ways in which privacy can be protected. I believe that one such method is for statistical experts to develop structured statistical methodologies and that these may lessen the burden of medical writers … eventually. What are structured statistical methodologies, and how might they help us in our work?

Cathal: So, I need to give you some background here before I get to the nub of your question. When anonymising documents, it can be difficult to know which data might be "identifying" and which data might not. For example, if you have 100 participants from the USA and only one from Ireland, it is fairly obvious that to protect the participant from Ireland you would need to redact "Ireland", but you could retain "USA" in your document without compromising the USA-based participants. But what about if you have 20 people from Ireland, and only five of them are female? Once you start including other identifiers, such as people's race, age, or ethnicity, it can become quite easy to make an individual stand out as highly identifiable. This is where things get complicated, and present challenges for medical writers. So, statisticians use quantitative risk to establish which rules should be applied to identifiers (age, gender, race, country, etc.) to protect patient privacy whilst maintaining data utility. It is generally thought good practice to group some values together rather than to redact all information. Let's say that you have participants from Ireland, England, Spain, and France. There may be low numbers of participants from each country, but if you group them together and label them as

"Europe", it may make the participants less identifiable while maintaining some geographic information. We tend to apply similar grouping with numeric values such as age, height, and weight. Instead of completely redacting these values, we group them in numeric bands. These are very simple examples of how we apply structured statistical methodologies to statistically anonymise data. There are more complex methodologies for more complex situations.

Sam: So now we can see how our statistician colleagues can do some of the work for us by anonymising datasets in this way. I imagine it can get very complicated to anonymise clinical trial datasets in this quantitative way. There is the risk of reverse-engineering data to identify an individual if data anonymisation is not done properly. How do you address that aspect when developing structured statistical methodologies?

Cathal: We address this by considering a measure called "k-anonymity". In short, k-anonymity is how many people share the same characteristics. For example, if you have a small group of three people, all from the same country, who are the same, race, gender, ethnicity, and of similar height and weight, then you have a 1 in 3 chance of "guessing" which person is which. So you take the number 1 and divide it by 3. This gives you a quantitative risk score of 0.33. The recommended threshold for public sharing is 0.09. What this number actually means is that each participant must share the same characteristics as at least 10 other people in the population. Note that I said population and not clinical trial. The population refers to the number of participants that you are using as your total for sharing.

Determining your population is often the first part of calculating your risk. The larger the population, the more likely it is that patients are going to share similar characteristics to other patients. There are several possibilities when it comes to determining an appropriate population for calculating the quantitative risk. Appropriate populations include:

Document Population

"Document population" indicates that you are only going to use the patients that appear in your document as the population for your risk calculations. This can be tough to achieve. The documents being shared will often concentrate on a subset of the patients that were in a clinical trial. This can mean that you have a very small population. When you have a very small population, then patients can be unique with just a few identifiers.

Clinical Trial Population

"Clinical trial population" is all the patients that took part in the trial, upon which the documents being shared are based. As mentioned earlier, not all patients in a clinical trial are discussed in the document. When we use the clinical trial population, we can use a larger population to calculate the risk of reidentification of patients. The larger the population the more likely that patients will share similar characteristics.

All Similar Sponsor Trial Population

If you wish to calculate the risk of reidentification, you could use an even larger population. Sponsors tend to carry out more than a single trial within their therapeutic area. If you group the data about these patients from all these trials together, you end up with a much larger population. As pointed out earlier, larger populations mean it is more likely that patients will share similar characteristics and reduce the risk of reidentification.

All Similar Trial Population

This is almost exactly like the "all similar sponsor trial population" except that you calculate risk based on trials from other sponsors as well as your own. In practice, this can be hard to do as you normally do not have access to the data from other sponsors.

Therapeutic Area Population

This is usually the largest population considered when calculating the risk of reidentification. This would be everyone in a geographic area within a therapeutic area. For example, all diabetics in the USA. As you can imagine, this would be a huge population. The difficulty is that I am unlikely to have access to this individual information for every single diabetic in the USA. However, there are documented techniques for population estimators.

Which population to choose often comes down to a sponsor perception as to what an attacker is likely to know. Will they know the name of the trial that the person they are attempting to reidentify took part in?

Sam: It is a relief for medical writers to know that statistical colleagues are right behind us in protecting clinical trial data within datasets. This will eventually lessen our burden and will mean that we will have less proactive anonymisation and less redaction to do in our clinical reports. This is, however, clearly an industry "work in progress", and the fact that EMA has set up the TAG for this purpose shows its importance. In your opinion, how long until this becomes the normal way of working i.e. statisticians routinely applying structured statistical methodologies to clinical trial datasets before we medical writers start to report the data?

Cathal: I think that we are rapidly changing our current procedures to prepare for public sharing. However, people are nervous when we have different agencies with different criteria that they need to meet. Ideally, we would see all agencies aligning their policies so we have one harmonised way of publicly sharing documents. Unfortunately, we are not there yet. Right now, Health Canada is the main driving force, and people are adapting to their public sharing policy, which is well-aligned with EMA's Policy 0070, although there are still some slight differences.

References

1. EMA Technical Anonymisation Group web page: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external- guidance-implementation-european-medicines-agency-policy-publication-clinical-data
2. EMA Policy 0070 and implementation guidance web page: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/ technical-anonymisation-group
3. CDISC Standard, STDM: https://www.cdisc.org/standards/foundational/sdtm
4. CDISC Standard ADaM: https://www.cdisc.org/standards/foundational/adam
5. CDISC home page: https://www.cdisc.org/new-to-cdisc