

# Harmonization as a key to ex-US and especially European Real World Data

I attended a panel discussion led by a number of life sciences real world data and clinical research leads. When asked about the way to succeed in Europe where it is trickier than the US here was their consensus advice.

**Harmonize across geographies** – Groups have been caught up in words like federated analytics. But harmonizing may include federation; it may also not use it. What is more important is to standardize/harmonize how studies are executed. So having related harmonized protocols that adapt to each country, manage the concept sets so they are in sync, standardize how contracting, IP, epidemiology, and publishing will be executed. There are already big efforts invested into harmonization such as the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) <https://www.ich.org/>.

**Adapt to local needs** – Although somewhat obvious there are local approaches in each country. Real World Data groups looking to expand into Europe may have to make adaptations to their IT infrastructure as an example. One organization put in place an AWS availability zone in Germany so they could do their analytics and comply with the local legal and cultural requirements for data to stay in Germany.

**Obtain good counsel for privacy and operating in those geographies** – While it is true that there are different and often more restrictive regulations in Europe it also isn't true that it's impossible to get research done. Many counsel with unfamiliarity will look at the risk and provide feedback that it is impossible to do what you may be considering because they are over reacting to the differences in the EU from the US. But with good and calm legal counsel regarding privacy many things that could be excluded are possible and may even be quite simple if the guidance is favorable to the approach being taken.

**Bring analytics to the data** – It is worth bringing the analytics to the data in many cases. When faced with tricky scenarios in the EU regarding data governance and sovereignty at large health systems tasked with protecting patient privacy it can be hard to just move data to a central store. But it is possible to overcome these challenges by getting good at moving the analytics to the data.

## **ChatGPT outbound requests from healthcare to Open AI are rapidly becoming adopted**

Yes this will include a comment about the adoption of AI in healthcare. Keep in mind that in general I'm not an evangelist of individual technologies such as cloud or of genAI. I prefer to see a business or clinical need and then have new platform tools add new ways to make the solutions better rather. In healthcare in particular I don't think the technology itself as being the solution since a heavy dose of governance and clinical workflow/change is a huge part of improvements.

GenAI and in particular capabilities in the LLMs such as chatGPT have demonstrated strong potential for solving hard issues in healthcare data. I won't talk about the specifics because there is too much chatter about every independent use case. What I would like to talk about is the remarkable shift in the approach for identifying patient data that has taken hold as a result of interest in using it for change.

I'd like to step back to a few years back when cloud computing was first introduced. The general feedback to it was that very few health systems wanted their data in the cloud. They didn't even want data warehouses in the cloud. The cloud was considered a risky space and one not suitable for PHI. PHI presents a big liability in that it can lead to a HIPAA breach which is a massive financial risk as well as a potentially bigger reputational risk.

As a result of this risk most health systems have had very tight technical controls or rules on identified patient data sets including clinical notes. The mantra that I have gotten used to is that because there is so much risk of a patient's identity being embedded in a clinical note that it is too risky to let notes leave the firewall of the health system. And for systems they do trust it is too risky to send out remote calls where identified data is shared. So historically almost all data processing steps such as natural language processing needed to be done by bringing the analytics to the data. This meant an architecture where every site needed to have a copy of some software package that could ingest the data and then output the needed classifications.

This is why I was surprised to learn that many health systems have agreed to directly interface data to ChatGPT for various applications. This means they are sending patient data to a remote server to process and return results. One person mentioned to me that they had given health systems a choice to either first run deidentification over the data and then send it to ChatGPT or send the data in a raw identified form. Deidentification in this case with free texts has to be quite conservative to make sure every possible action is taken to remove the PHI. As a result some data is lost through the redaction as an incidental impact.



What they found was that the majority of sites they were working on for this project involved extracting features when given the choice requested to send the identified data to chatGPT vs. first doing deID. What has apparently happened is that many health systems have adopted chatGPT through their IT department policies and have agreed to terms that enable a BAA and a HIPAA compatible set of agreements and services with Open AI or Microsoft Azure for the service. The following link is an example of the policies from Open AI regarding chatGPT and BAA agreements

(<https://help.openai.com/en/articles/8660679-how-can-i-get-a-business-associate-agreement-baa-with-openai>). In addition to this Microsoft has established their policies for how to collaborate.

### **So what's the big deal?**

The big deal is that we can stop obsessing about ways to avoid sending data to powerful LLM services that increasingly improve how we curate information out of free text notes written by clinical teams. This curated content is an ongoing pain point for efficiency in a number of areas of research or clinical decision support. Basically if a computer could make sense of clinician notes reliably then many of the tasks for a clinical research coordinator to read through records could be automated or get close to semi-automated. Furthermore, a decision support algorithm that relies on facts buried in notes can compile them and offer meaningful suggestions such as whether a patient should consider a genetic test or enroll in a trial.

For us at Graticule we can start including these services in client solutions for pharma when collaborating with health systems and be able to ride each wave of new updates to the LLM models. Since we don't have a long legacy of alternative approaches we can start fresh with a framework that will rapidly evolve to meet challenges and replace the art of the possible every year for the foreseeable future.

Interested in learning more?

**Reach out to us at [info@graticule.life](mailto:info@graticule.life)**