HDRUK's CO-CONNECT in a TRE is akin to storing landmines in a playground: you can convince yourself[1] it helps with greater goals, but is it responsible?

Data use can be dangerous by design, no matter how technically careful the implementation. It is not that HDR UK's CO-CONNECT project[2] is badly implemented as such, but rather that it cannot be made safe or trustworthy.

To use a Russian analogy, someone wishing to store landmines in a playground can loudly proclaim it will be made as safe as possible, and they may even believe it to be true. But at some point, some kid will go exploring – and they'll assume it won't be *their* responsibility…

In big letters on the front page of its website, CO-CONNECT says it is a "Curated and open analysis and research platform". This is a curious turn of phrase that goes to the heart of the flaws of the CO-CONNECT design.

If data is sensitive enough that it should not be disseminated, then it should not be touching this system. If data is so protected that it *can* be disseminated, then CO-CONNECT is an expensive and poor imitation of the many tools already available for that purpose. ('Not-invented-here' being endemic in HDR.)

Some of CO-CONNECT's public statements are unreliable

CO-CONNECT's website says it is for COVID, but its roadmap is to continue work now COPI has ended. The public (and medConfidential) therefore once again, with regard to an HDR project, find themselves in a place where CO-CONNECT's public statements are inconsistent with what CO-CONNECT is saying in private. We understand this compromised position will be rectified in due course, but it has not yet been done.

medConfidential can spend resources to determine the difference; members of the public should not have to.

CO-CONNECT is just one implementation of generic middleware

CO-CONNECT is a bit of middleware – it is the 'glue' between different systems that has to be configured on the HDR gateway end, and configured by the data custodian end, and only information configured can be passed.

There are many ways to show users a basic analysis, and this is one of them. The primary benefit to HDR appears to be that it is an HDR-native solution, reflecting the worst of "Not Invented Here" culture.

CO-CONNECT's choice of implementation brings its own risks.

---

[1] e.g. https://www.hrw.org/news/2020/02/27/questions-and-answers-new-us-landmine-policy
[2] https://co-connect.ac.uk/

Content of the data

CO-CONNECT is deliberately data agnostic, leaving all decisions to data custodians or controllers. It neither knows nor cares about the content of a dataset, treating variables as an abstraction handled by the metadata tool and the configuration.

As such, it is *not* designed to take account of the intricacies of complex national population datasets that should be accessed via TRE-only.

It is possible to get consent for this behaviour in a survey,[3] but that is not appropriate for administrative datasets of the entire population, like the GP records of a large swathe of Scotland.[4]

How CO-CONNECT is configured is entirely the responsibility of the gateway and the data custodian, and (bugs aside) CO-CONNECT will follow that configuration. Future upgrades, and the interactions of different current and future configurations options, also potentially raise significant risks that are complex to mitigate.

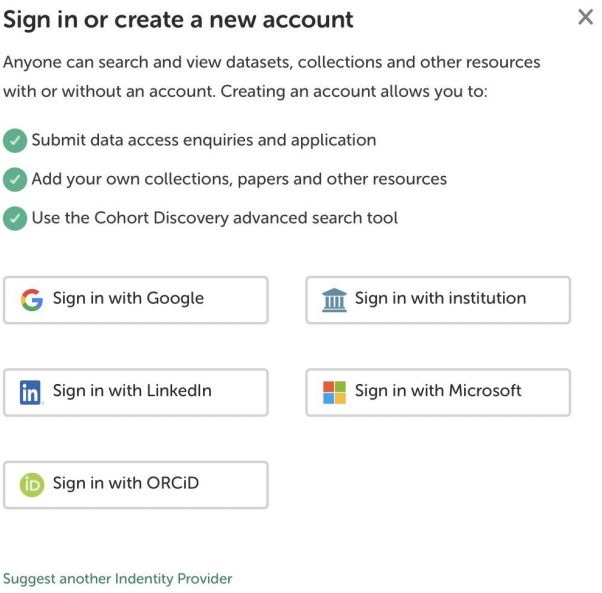
The CO-CONNECT metadata tool is clear that it can help with pseudonymising what the data custodian tells it are identifiers, but that is entirely the responsibility of the data custodian. Of course, identifiers that are not pseudonymised remain unaltered.

CO-CONNECT users come from anywhere; they are not just (accredited) researchers

Any count promoted on the front of a website is intended to go up, and HDR advertises its "2,329" "registered users on the front of its website.[5]

The login page shows a variety of "sign in" mechanisms that don't oblige users to be academics, which makes it impossible for a data owner to be consistent if they allow the CO-CONNECT / gateway integration for their datasets and also claim that the data has the security that comes from TRE-only.

It is entirely the responsibility of the data custodian to envisage every nuance that the tool could use, and to ensure none of them can cause problems – including the interactions between analyses by multiple "entirely different" researchers (which may be the same individual, logged in once via Google, once via their institution, and a third time via LinkedIn, etc).



[3] We can argue whether that is informed consent, and whether it is appropriate for a responsible data owner to use data unsafely even if they have permission, but surveys and small simple random sample datasets are not the primary concern when population scale datasets are also potentially covered.

[4] https://web.www.healthdatagateway.org/dataset/837a761e-a27a-45d9-ae61-bc3fc0b9f12a

[5] https://www.healthdatagateway.org/

We recognise that CO-CONNECT maintains an audit trail of what queries are run, which will be useful to show that extremely high risk queries run by LinkedIn or Google accounts were in fact run.

Trustworthiness, Quality, and Value?

The three pillars of the Office of Statistics Regulation's Code of Practice[6] are Trustworthiness, Quality, and Value.

Given HDR's goal that the gateway should have broad appeal, and the recognition that high risk datasets should be TRE-only, it is unclear how both of these characteristics can be assured either initially or over time.

And it is difficult to see how arbitrary queries responded to by persons (potentially) unknown can satisfy any of the three pillars. The cost of assessing this system will be nontrivial, and the cost of maintaining a *trustworthy* system as the gateway updates and evolves will be never-ending, complex, and impossible to understand externally.

CO-CONNECT concept is the brown M&Ms sitting in a bowl

The band Van Halen toured with a large set, needing a complex set of arrangements in each venue. Buried in the middle of the contract was a request for a bowl of M&Ms in the green room with all of the brown ones removed.[7] It was a shibboleth test. If the bowl was in the green room and had no brown M&Ms in it, then someone had paid attention to the details. If the bowl did have brown M&Ms in it, then *everything else* in the venue had to be carefully checked.

Approval for CO-CONNECT is a similar shibboleth test.

Due to the fundamental nature of the model, any data custodian or controller who says they use this tool *and* tries to reassure data subjects that their data is safe in their TRE should have every detail of their process carefully checked – simply because those two statements are incompatible.

medConfidential
October 2022

[6] https://osr.statisticsauthority.gov.uk/what-we-do/code-of-practice/
[7] https://en.wikipedia.org/wiki/Van_Halen#Contract_riders