Panel Q&A session from the Federated Analytics Webinar – 2pm, Thursday 9 May 2024

Watch the recording.

How intensive (or not) was the work each individual trust/hospital had to do (e.g. generating a common data model) in order to allow the federated analysis to occur accurately. Are you able to source data, verify, etc.?

- Andre: It is intensive. Typically in a project we would first define the, say 10, data elements that we need to answer the question. We would then "budget" 3-person months of the local data/IT staff to get those 10 data elements for that specific project. If a centre has done it for one project than typically it takes less time for the next project. The first project is the biggest hurdle.

- Stelios: I agree with Andre. Each hospital needs to source their data, verify it locally, and then format it according to a pre-specified common data model. This is one of the most time-consuming but also important parts of a federated learning project. As Andre mentioned, the first project within a network is usually the hardest to carry out, but once all participating hospitals become more familiar with the process, subsequent projects are easier and faster to carry out.

Really great work, Andre! You said, "this maybe isn't the best but it does work". Is there a "best" in your opinion?

- Andre: In terms of pure federated learning algorithms, we are impressed with what is done in Flower (https://flower.ai/). However they are focusing more on federating the algorithms rather than on how to create a federated learning network, including all the security, governance and supporting aspects.

We see different models in play around push or pull. We have found it "easier" when TREs pull the query down rather than them receiving queries directly.

- Andre: Correct – vantage6 is indeed also a pull system in which the hospitals are pulling a web page with instructions and if instructed to do that – for instance – pulls the analysis/docker to them. This allows for full control at the hospital and prevents the need for public IP/DMZ*.

*IP: Internet Protocol. DMZ: Demilitarised Zone.

Can you clarify on the open network point, does it require inbound ports to be opened, or only outbound?

- Stelios: Both inbound and outbound ports need to be open, as the communication between the local node and the server is two-way. In other words, a task / algorithm is sent from the server to the node (inbound), and then aggregated results are sent from the node back to the server (outbound). Port 443 is used as the default port for communication in vantage6 – this port is

usually open by default and information that travels through port 443 is encrypted and, therefore, secure while in transit.
- Andre: Correct, one needs to send data out but also needs to download data/dockers in vantage6. As Stelios says, this is through encrypted communication via 443 – the normal secure internet port.

In the "eyes-off" model of data access in fed learning, reliance on the data quality at each site is both vital and difficult to verify. Are there emerging sets of "standard" remote queries that can be run to sanity check quality across the whole cohort? Counts would be one, but are there others? (In physics simulations, conservation of energy was always the first thing to check, for example!).

- Stelios: I am not sure whether there are any standard algorithms for data quality assessment that have been implemented for use within vantage6 yet. However, bespoke algorithms can be designed depending on the needs of a specific project. For example, a federated script can be designed to check whether the dates in a dataset make sense (diagnosis date after date of birth, diagnosis date within study period, etc.), or whether the dataset includes any codes that we not specified in the common data model that is used for the project.
- Andre: We are indeed creating an "algorithm store" in which many standard algorithms will be available both classic distribution ones (e.g. means, standard deviation, per cent of missing data, etc.), imputation engines (for missing data), and outlier detection, etc. Also ones that allows one to compare a distribution from one hospital with another hospital to detect bias, different patient populations, etc. Note that we do not allow custom federated querying as with repeated queries a dishonest user can try to reconstruct the dataset of the hospital.

With the analysis coming to the data, how do you handle that the data varies over time? Do you end up with multiple versions of the same data?

- Andre: It depends on the project – we (in vantage6) do not take a position in that. Some projects want a snapshot of the data/persistent data, others want the data to be updated at every run. If the project wants different versions of data to be versioned/persistent this needs to be solved locally (by the data holder/station/hospital). The vantage6 tools are – by design – containers which are not persistent.

Can you give some pointers to the FOSS algorithm store and have you considered packaging the algorithms with RO-Crates and using metadata schemas such as croissant?

- Andre: The news item is here: https://vantage6.ai/news/new-algorithm-store-and-researcher-user-interface/. In terms of RO-crates and croissant, if I remember well, they are methods to make the algorithms itself FAIR. We are definitely working on that as this will allow building a FAIR catalogue of algorithms. Still very much work in progress. RO-crate is mostly used by (life sciences?) researchers – I have not yet seen it used by hospitals/real world data projects – which are the ones we are mostly involved in. But vantage6 is agnostic to the data/metadata scheme – it just orchestrates a federated learning network.

How does the commercial model work for vantage6?

- Andre: vantage6 itself is completely free and open source, so no license cost. Everyone can host it and support it. If a consortium wants to contract our spinout (www.medicaldataworks.nl) to support it, we charge a price for the hosting and a price per hour for support. To give you a feel: For a 12-month project of say five hospitals wanting to do a federated learning project we charge about €15k. [Note that this is a COI for me as I am involved in the spinout].

Thank you for the insightful webinar! It's fascinating to see the potential of federated analytics in healthcare research. My question pertains to the variability in data quality across different local sources and its potential impact on model metrics. How can you ensure that variations in data quality do not compromise the accuracy of the resulting model? For instance, is there a risk that instead of predicting cancer outcomes, the model might inadvertently identify the source of the testing entry due to disparities in data quality?

- Andre: Yes, this is a common concern. Not that this is also a concern when centralizing the data. Also in that case one may be concerned that one of the centres submitting data has a bias or a data quality issue that influences the end-result. The solution in a federated setting are also more or less the same as in a centralized setting (bias detection, imputation, filtering, excluding data or the whole cohort) only now they have to be done in a federated setting (so without "seeing" the individual data - but you can still share statistics!). In our project we are doing these things, but the end conclusion until now is always the same: data quality issues do not influence the end model a lot, it is almost always better to include ALL data (even if biased, low quality) as more data typically improves the end results.

Does vantage6 incorporate a multilingual terminology server to help with the international cohort building?

- Andre: No, a multilingual terminology server is typically needed to make the data FAIR. This is something that we push to the stations as it needs to be done locally. vantage6 assumes this has been done. (Note that we have such tools available because we obviously face these issues but they are not part of vantage6).

In the federated analyses I've seen there have been separate results for each database and really careful consideration has been given to each of these to account for local factors (varying data availability, population differences, different healthcare settings). In the learning approach it seems you end up with a single result. How and when do you consider the heterogeneity of the individual data sources – and whether some should be excluded from the analysis/analysed separately?

- A similar question was asked by the Chair during the live webinar; an answer was provided by Stelios during the live Q&A session.

For the description of the workflow, does it use TES/WES/RO-crates?

- An answer was provided by Andre during the live Q&A session.

You say that local collection and curation of data is still one the largest barriers. Please can you explain why it is such a problem?

- An answer was provided by Stelios during the live Q&A session.

What happens if the data owner does not have compute power? How does the analysis run?

- Stelios: This is usually not an issue when text-based data are being analysed, because the compute power needed for such analyses is rather low. However, considerable compute power might be needed to analyse images – in these cases, each participating centre needs to have enough power (e.g. a GPU) to analyse the local images. I don't think there is a way to analyse the data locally if the data owner does not have the necessary compute power. Please note that each centre only needs to contribute a fraction of the total compute power in each project, enough to analyse the local data.
- Andre: Agree with Stelios. At https://www.medicaldataworks.nl/security we have specified some common specs. [Note that this is a COI for me as I am involved in the spinout].

Why is Python to be removed as a dependency? Is that because it is going to become a container only application?

- Andre: Yes, the only thing Python does is basically to start up the dockers/containers. We feel this creates and extra dependency which is not necessary.

How can international researcher consortiums use this for studies? Is there tech support/consultation at all? Is vantage6 seen more favourably/eases during data sharing agreements?

- Stelios: It is my understanding that using federated learning eliminates the need for a data sharing agreement, as individual-level data are NOT being shared or sent outside the originating centre. However, a collaboration agreement and all the necessary ethics approvals still need to be in place before a federated learning project can commence. Using vantage6 and federated learning for a multi-centre project may be seen more favourably by an ethics committee when reviewing such projects, due to the privacy-preserving nature of this approach.
- Andre: Governance is definitely an issue – as Stelios also pointed out in his presentation. We are trying to help by not just publishing the source code of vantage6 as open-source, but the agreements are also open-access. You can find them on this website https://www.medicaldataworks.nl/governance. [Note that this is a COI for me as I am involved in the spinout].

How does vantage6 overlap/align with other hands-off federated approaches such as TRE-FX?

- Andre: I do not know enough about TRE-FX specifically to answer this question well. In general we know of many implementations of federated learning including (this list is not complete): vantage6, BranchKey, EHDEN, Linksight, Roseman Labs, Varian Learning Portal, International Data Spaces, Janssen Honeur/Feder8, DataSHIELD, NVIDIA Clara, FeatureCloud, Google Federated TensorFlow, PADME, Beacon, Hewlett-Packard, AusCAT, Scaleout, Flower, Philips, Mellody, Owkin/Substra. Ideally we are working towards merging some of these initiatives so that this list does not grow.

HDRUK
Health Data Research UK

How long did it take from initial discussions of atomCAT consortium to signing up agreements and then standardising the data?

- Stelios: For atomCAT2 (larger study involving 16 centres), it took a total of about two years from initial discussions – or rather, demonstrating proof-of-concept via atomCAT1 – to having all data ready for analysis in 16 centres. The most time-consuming aspects of the set-up process were legal/contractual-related: getting the collaboration agreement reviewed and signed by all centres and gaining the necessary ethics approvals in all participating centres. During this time however, we were able to carry out other tasks, such as finalising the statistical analysis plan, preparing and testing the federated algorithms/models, and helping centres familiarise themselves with the vantage6 infrastructure.

How much did you learn from federated data on top of what you would have learned from one data centre alone?

- Stelios: In the case of atomCAT2, I don't think that we would be able to learn much if we only analysed data from a single centre. For example, if we carried out the atomCAT2 analysis on LTHT data only, we would only be able to look at ~200 patients, which is a rather small cohort (even for a large regional hospital!). According to the sample size calculation, with this cohort size, we would only be able to assess the impact of one or two parameters on the outcomes we are exploring, and therefore the models would not be very informative.

  In addition, these local models would likely suffer from small sample size bias. By implementing a federated approach, we were able to analyse data from 1428 patients (during model training) and therefore, we could include up to 11 parameters in our models. This meant that the models developed are more robust, more informative, and more generalisable, since they were developed using data from patients treated across many different countries. Moreover, being able to externally validate the models using data from an additional 277 patients treated in two centres that did not participate in the model training, means that we can be more confident that the models are reproducible and representative of a varied patient population.

How do you monitor the quality of your models with reference to model decay over time?

- Stelios: We haven't set up any processes to monitor the quality of the models over time. However, I believe that such a process would not differ significantly from monitoring the quality of a centralised model. It would potentially be more time-consuming overall, as all centres would have to update their data at regular intervals, start their local nodes and the coordinating researcher would have to manually run all the models for re-training/updating.

How many deployments outside of your commercially supported networks do you have? The open-source version has issues in documentation, is there a community around this product?

- Andre: From our spinout [note my COI here] we are supporting about 15 federated learning projects at the moment (https://www.medicaldataworks.nl/customers). Most of these are non-commercial projects (Horizon Europe and similar types of funding). We have an additional five or so vantage6 projects which we host at the university, but we are moving these outside as the university is not equipped to provide the necessary support.

- Besides us, the Dutch Cancer Registry (a foundation) is a heavy user of vantage6 as they are doing federated learning across global registries. They are the ones managing the open-source community at the moment with two FTEs of dedicated engineers supporting (discord channel, roadmap, release management, etc.) In the coming years the management of this community will be migrated to the eScience center which is the Dutch national centre of expertise for research software (https://www.esciencecenter.nl/) so a more logical place to further build this open-source community than the cancer registry.
