LIV.INNO Data Science in Healthcare and Health Technologies Workshop:
Breakout Session Summary Report

Session: Synthetic data generation and validation standards
Chair: Dr. Debdeep Ghosal

Attendees:

- **Rob and Mark**: Early-stage PhD students in High-Energy Particle Physics.
    - Mainly dealing with Monte Carlo (MC) simulations using platforms such as Geant4, FLUKA, Root, and Pythia etc.
- **Nicolas**: Research collaboration manager with a background in Neuroscience, focusing on multi-institutional research efforts.
- **Carsten & Debdeep**: Dealing with beam instrumentation R&D works for accelerators, they use synthetic data and simulations as input beam distributions.

Main Discussion Points:

1. New Researchers and Grant Agencies

- Explored when grant agencies begin to 'take seriously' early-stage researchers (within 10 years post-PhD) seeking collaborative projects, what are the different pros and cons…
- Emphasis on aligning academic milestones, outcomes with fundable project ideas and immediate applicability.

2. Real vs. Synthetic Data

- Importance of understanding the *commercial value* of data as well as a balance between them: because although Real data are still signified as the 'gold standard', but synthetic data can rightly bridge the gap where *data sparsity* is a real issue. So, an inclusive combined approach should be way forward.
- Debate on making synthetic data viable and trustworthy for both academic and industry used cases, and while doing so, what would be the setting standards for synthetic data in clinical field? more like we have factors in physics as- statistical errors, reproducibility, clear data management practices etc.?

3. Academia vs. Industry Perspectives

- Trade-off between selling the *fundamentals* to appeal to industry versus the traditional academic approach.
- Discussed **disclosure record agreements** and IP protection when working across the academic–industry boundary.

- Need for clear deliverables and objectives in collaborations with industry (fast-paced, results-oriented targets).

4. Grants and Opportunities

- Identified available internal and external relevant grants:
    - National (UKRI)
    - UK/EU funding platforms (EPRC, Horizon Europe)
- Importance of focusing on innovation and cross-collaboration as core evaluation metrics.

5. Technical Aspects

- Compared traditional *statistical methods* vs. *deep learning* approaches: - while both of them have their own pros and cons..
  Classical statistical methods offer reliability and interpretability, making them ideal for smaller or well-understood datasets. Deep learning shines when dealing with complex, multi-modal data, but comes with higher computational and privacy considerations.
- Regarding the right balance between statistical fidelity, model utility, and privacy guarantees- ultimately, it's about balance! we must weigh all three of them to align with both clinical needs and ethical obligations.
- Critically examined consent and restriction implications when using publicly available data versus model training and validation. Because, while respecting the consent issue, one chucks out personal data and discard them, it could impact the training on the model if there is not enough data in the end and hence the validation will be limited.

Conclusions

- The session underscored the need for early-stage researchers to balance academic rigor with commercially viable innovation.
- Establishing robust standards for synthetic data can aid in gaining trust across disciplines and industries.
- Clear IP and disclosure agreements are critical for fruitful collaborations between academia and industry.
- Statistical versus deep learning approaches must be selected based on the project's priority: fidelity, utility, or privacy.
- Identifying and pursuing national and EU-level grants can boost long-term collaborative efforts, which should be the way forward.