Wellcome Data Science Ideathon Proposal - BASELINE

Background. Rapid vaccine development to novel diseases requires working with initially limited information, a lack of assays and outcome measures. We can use correlates of protection to help understand the spread of infectious disease, by identifying markers which relate to a protective immune response.

Early in an outbreak, we have limited information to help guide our pandemic response. However, we can be prepared for this situation by analysing vaccine responses of related diseases. This allows us to find robust correlates of immune response to established vaccines, which we can take forward to identify potentially at-risk populations.

The identification of these populations can help to guide policy recommendations on behaviour (such as limiting travel), precautions (such as masks) and vaccination - helping to optimise the use of vaccines, once developed, so they go to those needing it most, additionally guiding selection of populations in need of boosters.

Unfortunately, there are significant challenges with performing these analyses when time is limited. Studies are often siloed due to data governance legislation, which can limit our ability to combine studies. Furthermore, studies normally take a bespoke approach to data records which can add significant time when attempting to understand and prepare data.

We developed a two-pronged approach to help with these challenges. The first is a **novel method of rapidly analysing data structure and preparing them for challenge using a graph-theory based approach**. We have attempted to codify some of our subconscious analysis of datasets, which is normally performed by a manual search of the data.

The second approach is to use **high-dimensional reduction techniques and machine-learning clustering approaches which can be repeated on successive datasets**. This allows us to use many potential markers of various data types, and allows us to validate the techniques on other datasets. We can take this model forwards to help identify vulnerable populations early on in a future pandemic.

Platform Overview. Our graph-based data analysis and preparation approach provides a novel method using graph-theory which identifies the network structure of disparate data. It interprets the relationships between data fields by analysing the many-to-one relationships, then uses this to rank data fields using a random walk iterative method, by preference of joining. Graphs of the files connected by the data fields are created, allowing files to be joined by the preferred data field. The shortest path is then used to join together files needed for the data requested.

The BASELINE platform generates reference models of population composition from large-scale mixed numerical plus categorical data by learning a manifold to transform reference datasets using Uniform Manifold Approximation and Projection (UMAP). Model-based clustering approach Gaussian Mixture Model (GMM) is next applied to assign clusters within this low-dimensional space. By aligning clusters in the reference data with disease protection profiles (i.e. antibody readouts), the underlying quantitative and clinico-demographic features that correlate with disease protection.

The learnt UMAP+GMM model can be applied to new study data to rapidly assign new data into the reference-defined subpopulations. Additionally, new study-linked datasets (e.g. molecular readouts, immune phenotyping etc) can then also be stratified by model-assigned clusters to identify potential new correlates of protection for research.

BASELINE will provide both policy-guidance and research utility that highlights correlates of protected sub-populations and offers a "ready-to-deploy" tool for rapid identification of high-risk individuals to take preventative action in the event of future infectious disease outbreaks.

Work plan:
Months 1-6: Identify and engage with stakeholders and biostatisticians. Engage with key policy makers and MPs early to promote BASELINE's potential and define stakeholder requirements. Consult biostatisticians to regulate statistical interpretation.
Months 1-12: User-interface front end development: Comprehensive development and regular review of user experience.
Months 1-3: **Data Gathering:** We will use health data from HDRUK PIONEER. Allowing access to patients from a wide range of environmental,social economic groups, including ethnicities. This will be used for training,validation and optimisation of the models.We hope this approach will improve the representation of people from marginalised communities.
Months 3-6: Incorporation of Advanced ML: Concurrently, we will work on enhancing the platform's analytical capabilities. Parameters for Umap GMM will be optimised and tested.
Month 7: Testing and Iteration: With the implementation of both UX improvements and advanced ML, we will conduct thorough testing to ensure optimal performance.
Months 8-9: Integration of Diverse Data Types and Structures: Recognizing that different studies employ different kinds of data, we will focus on making our platform more versatile. This will involve developing algorithms capable of understanding and integrating diverse data types.
Month 10: Testing and Iteration: We will begin to introduce real world data, to help test its ability to integrate the different diverse data types which will test its ability to handle difficult datasets.
Month 11: Finalising the Platform and Preparing for Deployment: Upon successful testing, the platform will be finalised, and preparations will be made for deployment. This will include developing comprehensive user guides and training materials.
Month 12: Deployment and User Training: In the final month, the refined platform will be deployed.

Team member roles: Gokhan Tut: Team leader, stakeholder and public identification and liaison, optimise and test utility of tool. Duncan Murray: Development of graph based data exploration to incorporate multiple datastreams. Wayne Croft: Development of UMAP+GMM model, optimise tunable parameters. David Bone: UX implementation.

Collaboration with Biostatisticians: Our team will work closely with biostatisticians, especially who advise regulatory authorities. Expertise will guide the platform's development process and the interpretation of the findings. Regular workshops, training sessions, and continuous dialogue to overcome regulatory requirements and statistical best practices.

Working with Government/NHS Stakeholders: Transparent communication, active collaboration, and regulatory compliance will be central to our approach, making use of independent audits, workshops, training sessions, and evidence-based recommendations. This will help to build trust and ensure the broad acceptance of our platform and key to succeeding here will be to understand how this platform integrates with pandemic planning.

Adapting to New Diseases/Vaccines: We will ensure our modelling pipelines' adaptability and flexibility to accommodate new diseases and vaccines. The model generated to flu vaccine responses will be widely comparable to new vaccines developed as general immune response will also be measured for those having the new vaccine. The continual learning algorithms employed by the platform will enable iterative learning from new data, thus facilitating the platform's adaptability to evolving challenges.

Collaboration with LMICs: We aim to engage with researchers in Low- and Middle-Income Countries (LMICs) in a co-development approach to make the platform more responsive. Training workshops, dedicated support, and language localization features will be employed to democratise access to our platform in these regions.

Data Governance Statement: Our platform is committed to maintaining stringent data governance standards. We ensure strict adherence to privacy regulations like GDPR and HIPAA, utilise advanced encryption for data security, and enforce strict access controls. All code will be provided as a free open source licence to enable public involvement and use, with an inclusive code of conduct allowing for contributions.

Ethnicity, Diversity and Inclusiveness: The main team is made up of diverse levels of career stages allowing for a wide range of expertise to contribute to the platform development. We will also implement a **dynamic working environment** which will enhance individuals unique working styles and provide a **positive working culture**. This demonstrates **respect, values and diversity** within the team.

Expected Outcomes

1. A refined, user-friendly platform capable of incorporating a wide range of data types and structures to predict correlates of protection for infection and vaccine responses.
2. Establishment of a robust, ongoing collaboration with biostatisticians and government stakeholders and policy makers.
3. A platform capable of adapting to the evolving landscape of infectious diseases and vaccines.
4. Increased acceptance and trust of ML methods among key stakeholders and the public.
5. Wider accessibility and use of our platform in LMICs, contributing to global vaccine research.

In conclusion, our team is confident that the refinement and application of our platform will make a significant contribution to accelerating vaccine development, fostering global collaborations, and building trust in machine learning methods in vaccine research. We believe our work over the one-year grant period will yield vital advancements in the global fight against infectious diseases.

Appendix: Budget Template

Salaries:

Name	Justification	Period on project (months)	% time	Total
Gokhan Tut	Project lead	12	Full time	£50,000
Duncan Murray	Data linkage	12	Part time -10%	£10,000
Wayne Croft	Clustering and AI	12	Part time -25%	£15,000
David Bone	UI design	12	Part time- 25%	£8,000

Materials and consumables:

Description	Justification	Total

Equipment:

Description	Justification	Total
Hardware	Upgrading existing hardware	£5,000

Access charges:

Description	Justification	Total
Open AI API and other services	Use of API and other services provided by Open AI to help accelerate our work	£500
Cloud services to host and run platform	We would need to execute the final product on a cloud service to make it accessible by everyone.	£2,000
PIONEER	Access to health data	£5,000

Travel and subsistence:

Description	Justification	Total
Conferences	Would need to keep on top of the current AI trends and uses to make sure platform is modern and accessible	£1,000

Miscellaneous:

Description	Justification	Total
Consultant fees	When building and implementing a platform there may be a requirement from an expert like cloud computing or website design.	£3,500

Summary of costs requested:

Description	Total
Salaries	£83,000
Materials & Consumables	0
Equipment	£5,000
Access charges	£7,500
Travel and subsistence	£1,000
Miscellaneous – other	£3,500
Grand Total	£100,000